2019

Multivariate Statistical Methodologies used in In-vitro Raman Spectroscopy: Simulations and Applications for Drug and Nanoparticle Interactions

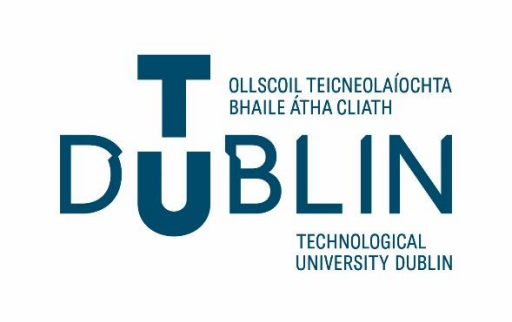
Mark Edward Keating
Technological University Dublin

Multivariate Statistical Methodologies used in *In-vitro* Raman Spectroscopy: Simulations and Applications for Drug and Nanoparticle Interactions

Mark Edward Keating

A thesis submitted for the Degree of Doctor of Philosophy

Supervisor
Prof. Hugh J. Byrne
Dr. Franck Bonnier



School of Physics/FOCAS Research Institute

Technological University Dublin

2019

Abstract

Raman spectroscopy is a growing technology in the fields of *in-vitro* drug and nanoparticle screening. The label free capability provided by vibrational spectroscopy, as well as the ability of the technique to probe the chemical nature of samples, makes it a good candidate for use in these fields. Crucial to the progress of these methods is the development and validation of robust and accurate multivariate statistical analysis protocols. In this thesis, both established and novel methods are examined using both real and simulated datasets. In particular, simulated datasets are used to validate and assess the accuracy of these methods in a spectroscopic setting. Firstly, partial least squares regression (PLSR) is examined using a simulated model based on real experimental data. This is applied to investigate the application of the algorithm to continuously varying data with known spectral perturbations introduced over a range of concentrations and responses. The results show that, while PLSR is valid for some dose ranges, sub-lethal, low concentrations and thus subtle spectral changes in the data may lead to difficulties in model construction. Multiple trends present in the data were also investigated and possible model error based on spectral bleedthrough in the regression coefficients RCs is explored. Principal component analysis (PCA) was also investigated using simulated datasets based on known changes in the data. Some of the limitations of PCA for data partitioning and trend analysis are overcome by a novel variant termed, 'seeded' PCA. 1st and 2nd derivative data is also explored for improvements in Raman spectral analysis using seeded PCA. Additionally, analytical methods used for Raman cellular imaging are also explored for nano applications, with two methods, classical least squares analysis

(CLSA) and a novel method spectral cross correlation analysis (SCCA) showing some improvements over current methodologies. Future work is also described pertaining to the use of a simulated cellular imaging dataset for validating data analysis protocols for spectral classification and *in-vitro* screening.

Declaration

I certify that this thesis, which I now submit for the degree of Doctor of Philosophy, is my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for postgraduate study by research of the Technological University Dublin and has not been submitted in whole or in part for an award in any other Institute or University.

The work reported on in this thesis conforms to the principles and requirements of the Institute's guidelines for Ethics in Research.

The Institute has permission to keep, to lend or to copy this thesis in whole or in part on condition that any such use of the material of the thesis be duly acknowledged.

Signature _____ Date _____

Mark Keating

Acknowledgements

I would like to firstly thank DIT, the school of Physics and the FOCAS research institute for accommodating me over the duration of this study. I would also like to thank my supervisors Franck and Hugh for putting up with me for the past four and half years of study and for all the help they have given along the way. I would also like to thank my fellow collegiates and lab-mates for the help and assistance over the duration of this PhD, both the old and the new. I would also like to thank my family who've stuck with me over the highs and lows, peaks and troughs of the Irish education system. I would also like to thank any friends, acquaintances and people whom I may have forgotten who have helped me through this experience.

Abbreviations

Abbreviation	**Full name**
AFM	Atomic Force Microscopy
CaF_2	Calcium Fluoride
CARS	Coherent Anti-Stokes Raman Spectroscopy
CCD	Charge Coupled Device
CLSA	Classical Least Squares Analysis
CLSM	Confocal Laser Scanning Microscope
dAMP	Deoxy Adenosine Monophosphate
dGDP	Deoxy Guanosine Monophosphate
DIC	Differential Interference Contrast
DMEM	Dulbecco's Modified Eagle Medium
DNA	Deoxyribonucleic Acid
EGFR	Epidermal Growth Factor Receptor
EM	Electron Microscopy
EPR	Enhanced Permeability and Retention
FCM	Fuzzy C-means
FITC	Fluorescein-5-Isothiocyanate
GCPQ	Quaternary Ammonium Palmitoyl Glycol Chitosan
HCA	Hierarchal Cluster Analysis
HPV	Human Papilloma Virus

IR	Infrared
KMCA	K-means Cluster Analysis
mRNA	Messenger RNA
NaCl	Sodium Chloride
OSCC	Oral Squamous Cell Carcinoma
PCA	Principal Component Analysis
PCL	Polycaprolactone
PEG	Polyethylene Glycol
PLGA	Poly Lactic-*co*-Glycolic Acid
PLSR	Partial Least Squares Regression
RES	Reticuloendothelial System
RNA	Ribonucleic Acid
rRNA	Ribosomal Ribonucleic Acid
RRS	Resonant Raman Spectroscopy
SCCA	Spectral Cross Correlation Analysis
SECARS	Surface Enhanced Coherent Anti-Stokes Raman Spectroscopy
SEM	Scanning Electron Microscopy
SePCA	Seeded Principal Component Analysis
SERS	Surface Enhanced Raman Spectroscopy
SESORS	Surface Enhanced Spatially Offset Raman Spectroscopy
SHG	Second Harmonic Generation
SORS	Spatially Offset Raman Spectroscopy

TEM	Transmission Electron Microscopy
TERS	Tip Enhanced Raman Spectroscopy
TPF	Two Photon Fluorescence
VCA	Vertex Component Analysis

Table of Contents

List of Figures

List of tables

Chapter 1: Introduction: Raman Spectroscopy In Nanomedicine and Drug Screening

1.1 Introduction

Currently, there is a drive and a need to develop new *in-vitro* technologies, which can be used for a range of applications, including screening for novel therapeutic strategies to evaluate the potential risks of nanomaterials as well as other toxic compounds. This follows new regulatory practices in both the EU and US (EU Directive-2010/63/EU and US Public Law 106-545, 2010, 106th Congress)[1,2], generally based on the 3 R's of Russell and Burch[3] to replace, reduce and refine the use of animals for scientific purposes.

In-vitro technologies are currently used for a wide range of applications such as testing the toxicological potential of certain compounds, identifying novel drug candidates as well as the assessment of novel nanoscale compounds[4–6]. This can involve high throughput evaluation of a battery of compounds and therapeutics, simultaneously allowing for rapid evaluation of these materials. Increasingly, these methods are being optimised to challenge the current convention of using animal models. In practice, novel methodologies should provide a bio-mimetic platform which can give the same end-point evaluation as an animal model at a fraction of the cost.

Emerging in the field of analytical science is the use of optical techniques to characterise biological processes. These include disease diagnostics e.g. various cancers[7], as well as novel approaches to the evaluation of therapeutics and nanomaterials *in-vitro*, *ex-vivo* and *in-vivo*[8–10]. For the most part, these optical

methods rely on the use of fluorescent probes, which, depending on the application, may be specifically designed dyes or fluorescent proteins which aim to provide an optical visualisation of these processes[11–13]. Newly developed nanoprobes can also be utilised for such purposes, whereby these nanomaterials may possess inherent optical properties which allow for a visual characterisation of their interaction, for example using microscopic techniques[14–16].

However, in other cases, nanomaterials must be tagged to allow for their visualisation e.g. with a fluorescent label[17]. While this is a viable technique, in some instances the fluorescent moiety may become labile, thus creating an ambiguity between particle and fluorescent probe[18]. Additionally, fluorescently tagging these materials may alter their properties, and size and charge of the nanomaterial are known to affect the interaction[19]. Asides from these issues, cost is also a factor, as the process of fluorescently labelling can become quite expensive and therefore there is scope for alternative strategies in the investigation of not only nano-bio interactions, but also other biological assays.

Increasingly, Raman spectroscopy has emerged as a versatile technique which has been used to study a number of different biological processes in a label free manner. Applications include disease diagnostics[7], cellular studies of drug interactions[20], as well as mapping nanobio interactions[21–24], to list but a few. The technique relies on the intrinsic chemical nature of a sample and therefore does not require any additional reagents other than the sample (as well as the test particle or chemical). Therefore, the technique circumvents the need for additional labels and probes to investigate a sample or process.

In such applications, there is often a large number of spectra acquired and thus interpretation soon becomes a problem. To tackle these issues, multivariate

statistical analysis is routinely applied to the data. Depending on the application, these methods may be used to distinguish between diseased and non-diseased[7,25], separate regions of a cell or tissue[26–28], extract out features which describe a process e.g. in toxicity[9] or drug interaction studies[26,8]. Fundamentally, these methods aim to classify the obtained spectra and thus provide a medium by which the information acquired can be grouped and interpreted.

Importantly, as with any method which aims to challenge the current paradigm, validation is a crucial concern. In most cases, Raman spectroscopic analyses are compared to 'gold standard' practices, be they in diagnostics (histopathological staining) or cellular studies (fluorescent dyes and labels). This verification is paramount and allows for a new method to be assessed against its established counterpart.

In this thesis, Raman spectroscopy is assessed as an *in-vitro* tool for the investigation of nanoparticle-cell and drug-cell interactions. Specifically, the multivariate methods which are applied to these problems are investigated. While there are a number of challenges faced by the biomedical vibrational spectroscopy community, such as sample preparation and instrument fidelity, which come under the umbrella term of spectral reproducibility, the multivariate statistical methodologies applied to these problems are not without their own caveats.

It is thus the aim of this thesis to investigate the application of multivariate protocols applied in Raman spectroscopy and explore some of the potential pit falls associated with their application. The studies utilise simulated datasets, based on real experimental data, which contain known spectral perturbations, such that the intricacies of these multivariate statistical methods can be explored and the validity, sensitivity and limits of detection can be evaluated.

This work also lays down the core foundations of a supervised data mining approach in spectral cross correlation analysis, providing a novel approach to tracking nanoparticles in cells using Raman spectroscopy. As validation is an important concern, the method was compared to other data mining approaches such as classical least squares analysis, and further comparisons were made to K-Means Cluster Analysis, employed in the original work by Dorney et al[24]. While improvements in specificity were made in the identification of the intracellular nanoparticles, it was still not possible to determine whether all spectra in the dataset were correctly identified as containing nanoparticles or not, and thus a more complex cellular simulation was developed to investigate this issue.

This thesis is therefore laid out as follows. The background section is split into two chapters; chapter 2 is adapted from a review paper published in the journal: Nanomedicine, 8(8), 1375 – 1391(2013)[29], entitled; "Raman spectroscopy in nanomedicine: current status and future perspectives". The main focus here is to look at where Raman spectroscopy has been applied in a nanomedical context, focusing on some of the variants (namely SERS, TERS and CARS) and also investigating where spontaneous Raman has been applied. Thus, this section aims to give the reader an introduction to the nanomedical field, specifically from a Raman perspective, but also introduce some important concepts in the fields of nano science and nano biology.

Chapter 3 aims to describe the role of multivariate statistical methods and their application to Raman spectroscopy in general, but also where these methods have been used in the context of Raman cellular imaging, as well as some of their applications in exploring nano-bio and drug interactions. Some of the benefits and

short comings of these methods will be discussed and the concept of supervised approaches in Raman spectral data mining will be introduced.

Chapter 4 is a reproduction of the journal publication in Analyst, 140, 2482-2492 (2015)[30], and outlines the development of a simulated dataset to assess the validity of the partial least squares regression (PLSR) algorithm used in biomedical Raman spectroscopy. Based on the experimental results of Nawaz et al[26], a simulated dataset is generated with known spectral perturbations related to both concentration of chemotherapeutic agent and the resultant cytotoxic response *in-vitro*. Both lethal and sub-lethal dose ranges are explored with the aim of testing the limits of the PLSR algorithm and also identifying some of the potential pit falls of applying this method in Raman spectroscopy.

Chapter 5 details further investigations of multivariate statistical methodologies, namely principal component analysis (PCA), applied to biomedical Raman spectroscopy, using simulated data. Furthermore, a novel variant of the PCA algorithm is developed, termed seeded PCA (SePCA) and is shown to be superior to the standard algorithm for handling continuously varying data. Further insights are also garnered on the use of 1st and 2nd derivative spectra and the impact this mathematical transformation has on the ability of the algorithm to separate and describe the spectral origin of differentiation of spectral datasets.

Chapter 6 describes the development and application of a novel supervised data mining approach, spectral cross correlation analysis (SCCA) applied to Raman spectral data containing polystyrene nanoparticles, as well as specifically designed simulated datasets, based on a publication in Analyst, 137, 5792-5802 (2012). The approach is compared to a supervised and unsupervised

method in classical least squares analysis (CLSA). SCCA is also demonstrated as a method to identify other biochemical distributions in the cell, namely lipid and RNA distributions

Chapter 7 outlines the final discussion and conclusions drawn from this thesis, highlighting the importance of multivariate statistical analysis in an *in-vitro* Raman spectral platform for *in-vitro* screening technologies, with a particular focus on advancing data simulation in this context.

1.2 References

1 THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION, *Off. J. Eur. Union*, 2010, 33–79.

2 U. S. Congress, 2001, 2721–2725.

3 W. Russell, R. Burch and C. Hume, *The principles of humane experimental technique*, Methuen, London, 1959.

4 M. A. Maher, P. C. Naha, S. P. Mukherjee and H. J. Byrne, *Toxicol. In Vitr.*, 2014, **28**, 1449–60.

5 M. Davoren, E. Herzog, A. Casey, B. Cottineau, G. Chambers, H. J. Byrne and F. M. Lyng, *Toxicol. In-Vitro*, 2007, **21**, 438–48.

6 S. P. Mukherjee, F. M. Lyng, A. Garcia, M. Davoren, and H. J.Byrne. *Toxocology Appl. Pharmacol.*, 2010, **248**, 259–68.

7 F. M. Lyng, E. O. Faoláin, J. Conroy, A. D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan and H. J. Byrne, *Exp. Mol. Pathol.*, 2007, **82**, 121–9.

8 H. Nawaz, F. Bonnier, A. D. Meade, F. M. Lyng and H. J. Byrne, *Analyst*, 2011, **136**, 2450–63.

9 P. Knief, C. Clarke, E. Herzog, M. Davoren, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2009, **134**, 1182–91.

10 Z. Farhane, F. Bonnier, A. Casey and H. J. Byrne, *Analyst*, 2015, **140**, 4212–4223.

11 P. Sandin, L. W. Fitzpatrick, J. C. Simpson and K. A. Dawson, *ACS Nano*, 2012, **6**, 1513–21.

12 F. Bonnier, M. Keating, T. Wróbel, K. Majzner, M. Baranska, A. Garcia, A. Blanco and H. J. Byrne, *Toxicol. Vitr.*, 2014, **29**, 124–131.

13 S. S. Kelkar and T. M. Reineke, *Bioconjug. Chem.*, 2011, **22**, 1879–903.

14 X. Zheng, J. Tian, L. Weng, L. Wu, Q. Jin, J. Zhao and L. Wang, *Nanotechnology*, 2012, **23**, 055102.

15 S. J. Shin, J. R. Beech and K. A. Kelly, *Integr. Biol. (Camb).*, 2012.

16 J. Kneipp, H. Kneipp, A. Rajadurai, R. W. Redmond and K. Kneipp, *J. Raman Spectrosc.*, 2009, **40**, 1–5.

17 M. Pellach, J. Goldshtein, O. Ziv-polat and S. Margel, *"Journal Photochem. Photobiol. A Chem.*, 2012, **228**, 60–67.

18 T. Tenuta, M. P. Monopoli, J. Kim, A. Salvati, K. A. Dawson, P. Sandin and I. Lynch, *PLoS One*, 2011, **6**, e25556.

19 A. Verma and F. Stellacci, *Small*, 2010, **6**, 12–21.

20 K. Klein, A. M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F. Jamitzky, G. Morfill, R. W. Stark and J. Schlegel, *Biophys. J.*, 2012, **102**, 360–8.

21 H.-J. van Manen, Y. M. Kraan, D. Roos and C. Otto, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 10159–64.

22 T. Chernenko, R. R. Sawant, M. Miljkovic, L. Quintero, M. Diem and V. Torchilin, *Mol. Pharm.*, 2012, **9**, 930–6.

23 T. Chernenko, C. Matthäus, L. Milane, L. Quintero, M. Amiji and M. Diem, *ACS Nano*, 2009, **3**, 3552–9.

24 J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers and H. J. Byrne, *Analyst*, 2012, **137**, 1111–9.

25 A. Lattermann, C. Matthäus, N. Bergner, C. Beleites, B. F. Romeike, C. Krafft, B. R. Brehm and J. Popp, *J. Biophotonics*, 2013, **6**, 110–21.

26 H. Nawaz, F. Bonnier, P. Knief, O. Howe, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2010, **135**, 3070–6.

27 C. Matthäus, T. Chernenko, J. a Newmark, C. M. Warner and M. Diem, *Biophys. J.*, 2007, **93**, 668–73.

28 M. Hedegaard, C. Matthäus, S. Hassing, C. Krafft, M. Diem and J. Popp, *Theor. Chem. Acc.*, 2011, **130**, 1249–1260.

29 M. E. Keating and H. J. Byrne, *Nanomedicine (Lond).*, 2013, **8**, 1335–51.

30 M. E. Keating, H. Nawaz, F. Bonnier and H. J. Byrne, *Analyst*, 2015, **140**, 2482–2492.

Chapter 2: Raman spectroscopy in nanomedicine: current status and future perspectives

The following review paper was written by the primary author Mark E. Keating, while Hugh J. Byrne, as supervisor, was primarily responsible for editing and refining of the text. The format is that of the journal publication, but section and figure numbers have been adapted to the format of this thesis.

Keating ME, Byrne HJ. Raman spectroscopy in nanomedicine: current status and future perspective. Nanomedicine (Lond). 2013 Aug;8(8):1335-51. doi: 10.2217/nnm.13.108.

2.1 Abstract:

Raman spectroscopy is a branch of vibration spectroscopy which is capable of probing the chemical composition of materials. Recent advances in Raman microscopy have added significantly to the range of applications which now extend from medical diagnostics to exploring interfaces between biological organisms and nanomaterials. In this review, Raman is introduced in a general context, highlighting some of the areas in which the technique has found success in the past, as well as some of the potential benefits it offers over other analytical modalities. The subset of Raman techniques which specifically probe the nanoscale, namely Surface Enhanced and Tip Enhanced Raman Spectroscopy, will be described and specific applications relevant to nanomedical applications will be reviewed. Progress in the use of traditional label-free Raman applied to investigation of nanoscale interactions will be described, and recent developments in Coherent Anti-Stokes Raman Scattering will be explored, particularly applications to biomedical and nanomedical fields.

2.2 Introduction

Nanomedicine can be defined as the medical applications of nanotechnology[1], ranging from the use of nanomaterials in regenerative medicine, drug delivery strategies, medical diagnostics and therapeutics and including potential negative impacts of nanomaterials to human health, commonly encompassed under the term Nanotoxicology. In the context of this review article, nanomedicine is viewed from the perspective of how Raman spectroscopy (and its variants) can be used in the assessment of the beneficial as well as the potential negative impacts of Nanomaterials on human health. Nanomaterials have already found uses in a wide range of applications, including anti-microbial paint coatings[2], textile finishing[3], and novel applications in the electronics industry[4]. Notably, biomedical applications are rapidly emerging, ranging from nanoparticle coated stents for angioplasty[5], contrast agents for diagnostic imaging[6,7] and also potential drug and gene delivery vehicles[8–10]. These applications are largely dependent on the particular characteristics which nanomaterials and nanoparticles possess. These include properties such as increased surface to mass ratio which in turn results in an increase in surface reactivity, while novel optical properties associated with some classes of nanoparticles are important for applications in theranostic imaging and subsequent monitoring of drug delivery. However, whilst these technologies show promise, it is important to be able to visualise how the materials behave *in situ*, and particularly in the biological context, to be able to characterise their interactions and toxicological effects, be they *in-vitro* or *in-vivo*. While it has been highlighted that comprehensive characterisation of the physico-chemical properties of nanoparticles is imperative, changes to these properties,

such as aggregation state and effective surface chemistry, can play a critical role in their modes of interaction and action[11]. Equally, to understand the modes of action and optimise efficacies, monitoring and understanding changes to the biological environment is critical, not only on a cellular level but also when considering the systemic responses.

Considering the system as a whole, one must be able to track a particle or material from initial exposure or administration through to the site of action and on to assimilation, degradation or excretion. At each step in this process, one must be able to access and visualise the efficacy by which the particles can overcome certain barriers to successful administration. These can vary from the route of exposure, assessing whether the particle causes toxicity, particle retention (e.g. via the enhanced permeability and retention (EPR) effect), or removal for circulation via uptake by the reticuloendothelial system (RES), accumulation of the nanoparticles over time, non-specific interactions, the efficacy with which the particle reaches its desired location etc..

Ideally, what is required is a method which can successfully characterise these processes, firstly in fundamental *in-vitro* cytological and *ex-vivo* histological studies and ultimately in more realistic *in-vivo* applications. This method should be capable of identifying the particle or material of interest while simultaneously being able to access the surrounding environment while measuring the efficacy of the probe or nanocarrier and/or the physiological response of the organism.

There exists a large range of analytical methods which can be used in the classification and characterisation of nanomaterials. These include scanning and transmission electron microscopy (SEM and TEM), atomic force microscopy

(AFM), other label free optical methods such as differential interference contrast (DIC) and dark field microscopy and fluorescent microscopy methods based on intrinsic nanoparticle or external label fluorescence, to name but a few. However, these methods are not without certain drawbacks which limit to some extent their applicability and effectiveness.

Firstly, both AFM and SEM can be considered as primarily surface sensitive techniques, while, when TEM is coupled with serial sectioning and ultra-microtomy, it has been used for 3D reconstructions and tomography[12,13]. However, these processes are time consuming, costly and laborious. In addition, EM requires a particle to have contrasting electron density compared to its environment to allow for a particle to be visualised, which renders it ineffective for many "softer" polymeric nanoparticles. EM does not allow live cells to be imaged and, as it requires extensive sample processing, it provides only a limited scope for rapid or routine investigation of nanomaterials *in-vitro*. What EM and AFM do provide is the capability of imaging beyond the optical diffraction limit. More recently developed optical based methods, so-called super resolution microscopy, have become available that allow for imaging beyond this limit[14–16]. However, their use has been limited in the field of nanomedical sciences as of yet.

In contrast, standard fluorescent based microscopy has been used extensively in nanoparticle studies[16–20]. Confocal Laser Scanning (fluorescence) Microscopy (CLSM) has become a standard in the toolbox of techniques for *in - vitro* cytommetry [21]. Although the technique is limited in resolution to hundreds of nanometers, it can potentially detect fluorescence emission from, and therefore the location of, individual nanoparticles. Penetration depths *in-vivo* can be extended through two photon excitation techniques and/or NIR fluorophores[22,23].

In the visible region, a range of fluorescent assays and labels are commercially available to probe a range of physiological processes *in-vitro,* such as lyso and mitotracker used for labelling lysosomes and mitochondria[24]. Intrinsically fluorescent nanoparticles such as inorganic semiconductor quantum dots have been developed for similar applications[25] and surface functionalisation of these types of materials has contributed to understanding the dependence of uptake and intracellular trafficking on surface chemistry[26]. Many similar studies have been performed with fluorescently labelled nanoparticles[27,28] which are commercially available in a range of sizes and surface functionalities. However, not all nanoparticles can be easily fluorescently labelled. Furthermore, it is not clear that the transport mechanisms of smaller nanoparticles, fluorescently labelled with anionic moieties, are the same as their unlabelled counterparts[29]. Critically, there have been reports that labelled nanoparticles can release the dye into the surrounding biological environment, and so the distribution of fluorescence within the cell does not necessarily represent the presence or subcellular distribution of the nanoparticles[30–32]. Other label free optical microscopy techniques are also limited by the type of particle which can be visualised i.e. only metal based particles are effective for dark field and DIC microscopy[33].

Raman spectroscopy has been proposed as a method for monitoring nanomaterials in biological systems, as it potentially provides a label free, non-invasive probe of the nanoparticle itself, the local environment and the physiology of the organism[34]. Over the past decade, Raman spectroscopy has been applied to a range of biomedical areas, including cancer diagnostics[35], toxicity studies[36], atherosclerosis[37] and investigation of skin[38,39]. Importantly, what Raman provides is not just a method for differentiation between a diseased and non-diseased state,

it is based on characterisation of the (bio) chemical nature of a sample, based on the characteristic vibrations of the molecular bonds of the constituent components. Raman is a form of vibrational spectroscopy, which in itself is a subset of a more general umbrella term of spectroscopy. The vibrations are characteristic of the molecular structure and, in polyatomic molecules, give rise to a spectroscopic "fingerprint". The spectrum of vibrational energies can thus be employed to characterise a molecular structure, or changes to it due to the local environment or external factors. The Raman spectrum is thus a truly label free signature of the nanoparticle. Vibrational energies typically fall in the mid Infrared (IR) region of the electromagnetic spectrum and are quite commonly probed using IR absorption spectroscopy. Raman in many ways can be viewed as a complementary technique to IR spectroscopy; whereas IR involves absorption of radiation, Raman is an inelastic scattering technique whereby the incident radiation couples with the vibrating polarisation of the molecule and thus generates or annihilates a vibration. For a vibration to be active in IR spectroscopy, a change in dipole is required, whereas to be Raman active, a change in polarisability is required. As a rule of thumb, vibrations of asymmetric, polar bonds tend to be strong in IR spectra, whereas Raman is particularly suitable as a probe of symmetric, nonpolar groups. Importantly, this results in the O-H bonds of water being strong absorbers in IR spectroscopy, whereas they are relatively weak Raman scatterers. This allows for samples to be investigated in an aqueous environment and thus the technique of Raman spectroscopy more readily lends itself to live cell *in-vitro*[40] or *in-vivo*[41] measurement. As the vibrational spectrum is measured as a frequency (or energy) shift from that of the incident radiation, Raman spectroscopy can be performed across the UV, visible

or near infrared spectral regions, and thus can benefit from the technologies available and advances made for confocal optical microscopy.

A number of variants which are based around the physical principal of Raman spectroscopy exist. Spontaneous Raman can take the form of Stokes Raman scattering and anti-Stokes Raman scattering, the former resulting from the creation of a vibration in a material, characterised by a decrease in the incident photon energy (frequency), the latter from the annihilation of vibration, characterised by an increase in the incident photon energy. If the incident radiation is resonant with an electronic absorption of the analyte, the Raman signal can be resonantly enhanced by several orders of magnitude. The use of Resonant Raman Spectroscopy (RRS) in biomedical systems has been limited, however, due to associated photochemical degradation phenomena and the generation of fluorescence which can swamp the Raman signal of the overall sample.

Other variants of these two techniques with increased sensitivities for more molecularly specific characterisation have been developed. These include resonant Raman spectroscopy, coherent anti Stokes Raman spectroscopy (CARS), tip enhanced Raman spectroscopy (TERS) and surface enhanced Raman spectroscopy (SERS). The majority of these techniques have been applied to nanomedical applications; however, two of these methods deal inherently with the nanoscale, namely TERS and SERS. Although Raman is fundamentally an optical technique and is thus similar to confocal optical microscopy, limited to spatial resolution of the order of hundreds of nanometres, nanometre resolution can be obtained through localised enhancement processes. This localised

enhancement led to the initial interest in the prospect of the use of Raman spectroscopy to probe the specific environment of the nanoparticle.

This article will outline the applications of the various Raman spectroscopy based technique in the broad area of Nanomedicine. As they are nano-specific, the use of SERS and TERS techniques will be presented initially, while the increasing interest in the use of truly label free spontaneous Raman and Coherent Anti-Stokes Raman Spectroscopy (CARS) in nanomedical applications will then be explored. In Raman spectroscopy, the sensitivity, spatial resolution and penetration depth and required scan rates depend on technique employed, resonance conditions and even the instrumental set-up (microscope objective, grating, laser power). In the respective section describing each modality, examples of the state of the art in nanomedical applications are provided. The future perspectives attempts to address routes beyond the current state of the art. A more detailed description of the historical origin and basic principles of the Raman scattering process can be found in numerous excellent text books[42–46] and review articles[47–49]. A comparison of Raman and IR spectroscopies for biomedical applications can be found in [50].

2.3 SERS

The phenomenon of surface enhanced Raman spectroscopy was described as early as 1974[51,52], and is understood to arise from a localised increase in the coupling between the electromagnetic field of the incident radiation and the polarisation of the analyte in the presence of optically induced surface plasmons on a metal surface. Increases of Raman intensities as high as 10^{10} have been reported[53], although the spatial range of enhancement is only of the order of tens

of nanometers. The enhancement process can be achieved using a number of substrates including roughened metallic surfaces, structured metal arrays and specially imprinted surfaces.

Notably, the SERS effect can be induced through the use of metallic nanoparticles and nano colloid aggregates. SERS is a direct enhancement of the Raman signal and in the case of nanoparticles this occurs in the immediately surrounding local vicinity. The true principal that governs SERS enhancement is not fully understood, although the effect has largely been attributed to an electronic enhancement due to local fields generated by surface plasmon resonances at the metal surface. Alternatively, the enhancement has been attributed to a charge transfer process between the analyte and the surface, although it is probable that the processes act in tandem[54]. The technique of SERS in a biomedical context is reviewed in greater detail in the following papers[55–57].

Nanoparticles and aggregates which are used for SERS enhancement typically consist of a metallic nanoparticle, most commonly gold and silver. Quite often, these particles are subsequently modified via surface functionalisation which can include targeting moieties designed for specific applications, especially as nanosensors. The particle may also be labelled with a Raman reporter moiety which allows for identification of the particle in the biological milieu. Using these particles, it has thus been possible to apply SERS to a number of biological scenarios, which include diagnostic studies *in-vitro*, *ex-vivo*[58,59] and *in-vivo*[60,61] , novel bio assays[62–64] as well as cellular studies.

SERS has been proposed as a method for understanding how nanomaterials behave in a cellular environment, important in the study of the fundamental interactions of nanoparticles in the context of toxicology, drug

delivery or contrast agents for diagnostics. In 2009, Kneipp et al. proposed that by using SERS it would be possible to probe the chemical nature of the subcellular environment and the intracellular distribution of biomolecules. This work was extended by incorporation of Raman reporters which allowed for localisation of the SERS probe within the cell, leading to chemical probing of sub cellular nanostructures[65–68]. For example, in 2010, the group showed how a SERS nanosensor was capable of investigating pH changes in a cell throughout the stages of the endocytic pathway of the nanoparticle probe. The study was based on changes in the pH of the local environment in different cellular organelles which can be monitored via changes in the pH sensitive nanoprobe over time[69].

Other cellular studies have also investigated the possible use of SERS in the investigation of cell surface receptors associated with cancer. In one such study, Kong et al 2012 used organometallic SERS active nanoparticles which were targeted to live cells expressing the EGFR (epidermal growth factor receptor). The SERS nanoparticles were shown to be capable of specific targeting to the cell surface and offered increased sensitivity in comparison to other imaging modalities[70]. Figure 2.1 a-e, shows oral squamous cell carcinoma (OSCC) cells expressing the epidermal growth factor receptor (EGFR), c and e show the SERS image generated by CO at 2030cm^{-1} and protein at 1600cm^{-1} respectively. The targeting is verified in Figure 2.1 f – j in a non-EGFR expressing cell line SKOV3 (ovarian carcinoma) and in Figure 2.1 k-p by blockage of the EGFR using an EGFR antibody.

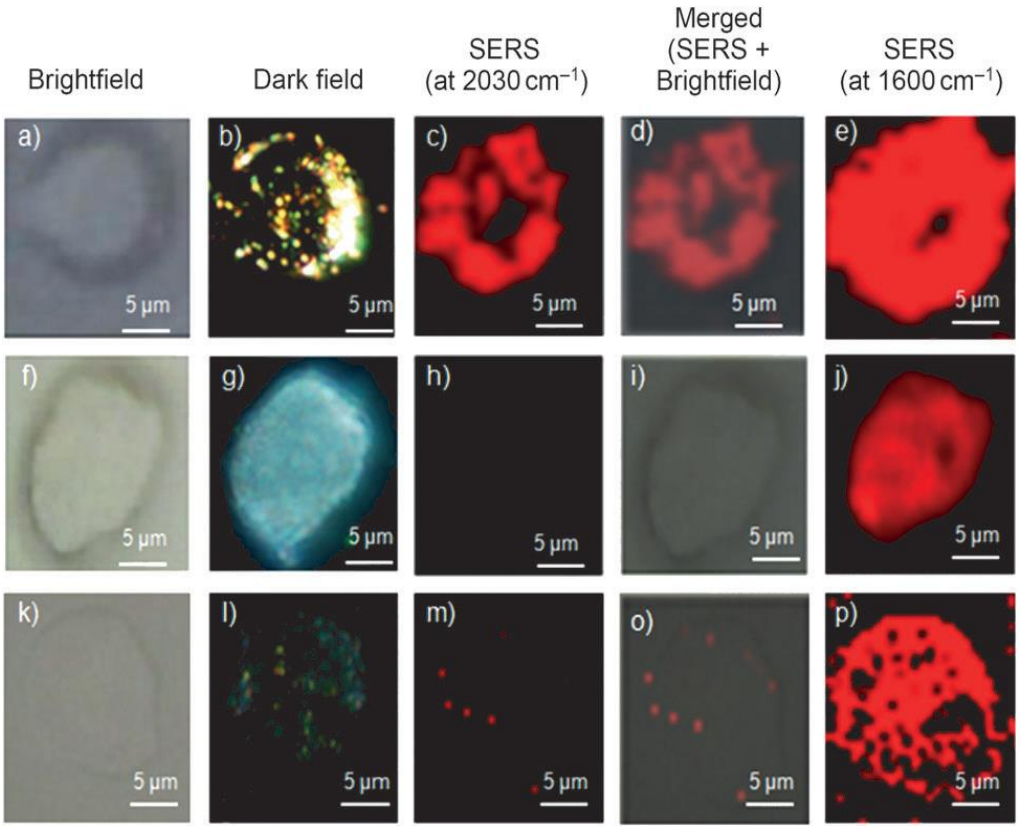


Figure 2.1 shows (a, f, k) the brightfield image, (b, g, l) the darkfield image of the nanoparticles, c, h, m) the SERS image of CO at 2030cm^{-1}, (d, I, o) merged SERS and brightfield, and (e, j, p) the SERS image generated using the protein band at 1600cm^{1}. a – e shows OSCC cells, f – g SKOV3 cells not expressing EGFR and k – p OSCC cells treated with anti-EGFR. Reproduced from [70]

Another demonstrated application of SERS in is the analysis of human serum. Lin et al., in 2011, demonstrated the power of SERS coupled with multivariate analysis to distinguish in a non-invasive way between patients previously diagnosed with colorectal cancer and control patients with 100% diagnostic sensitivity and specificity[71].

In-vivo SERS is also possible, and has been demonstrated as a potential labelling method for a number of applications. SERS has been used *in-vivo* to investigate how enhancement of the Raman signal can be used as a method for tumour detection. Qian et al. showed how EGFR targeting PEGylated gold

nanoparticles labelled with a Raman reporter where capable of >200 times greater signal generation in the infrared compared to that of near infrared fluorescent quantum dots, which allowed for the possible identification of small tumours at penetration depths of ~ 1-2cm[60]. Other *in-vivo* applications of SERS have also been explored, including an *in-vivo* study of inflammation in mice[72], demonstrating improvements over fluorescent based methods. SERS has also been shown to be capable of single molecule detection *in-vitro,* a sensitivity which sets it apart from spontaneous Raman spectroscopy[73].

More complex Raman based investigations have also taken advantage of the surface enhancement process. Techniques such as deep penetrating spatially offset Raman (SORS) have been combined with nanoparticle based SERS in SESORS[7475]. In brief, in the SORS technique, introduced in a paper by Matousek *et al,* the Raman spectra are collected at positions spatially offset from the point of incidence of the probe laser beam. Rather than using microscopic objectives for delivery and collection, fibre probes are used. By moving the collection point away from the probe launch site, contributions from the surface Raman photons are diminished and those of Raman photons from deeper within the sample are increased. Using multivariate statistical methods, it is possible to reconstruct spectra from the different layers with a much greater depth of penetration than a traditional confocal microscopy setup[76]. Depth sensitivities of up to several millimeters are now achievable and examples of emerging applications include non-invasive diagnosis of bone disease, cancer and monitoring of glucose levels[77] SESORS uses this same principal, taking advantage of the surface enhancement of the Raman signal from metallic nanoparticles embedded within the sample. In a recent publication by Xie et al, SESORS was used to identify bisphosphonate-

functionalized nanotags on bone through 20mm of porcine tissue[78]. This study highlights the increasing potential for *in-vivo* applications which SORS and SESORS may have, in the field of nanomedicine.

SERS has enjoyed increasing popularity over the past decade, particularly since the emergence of as increasing range of nanoprobes. However, the uptake rates and mechanisms as well as the subsequent trafficking may be specific to the nanoparticle type, size and surface chemistry. Most SERS probes are specifically designed for a target application and so are labels themselves for the SERS signal. Furthermore, the molecular specificity of the surface enhancement process is not well understood. Therefore, a truly label free method for generic monitoring and characterising the cellular uptake and subcellular localisation of nanoparticles in general is still required.

TERS another method for generating enhancement of the Raman signal. Like nanoparticle based SERS, this method is also based on probing of the inherent nanoscale environment of the sample in close proximity to a nanoprobe and will therefore be discussed.

2.4 TERS

Tip Enhanced Raman spectroscopy, or TERS, is a method which combines Raman spectroscopy and scanning probe microscopic techniques such as AFM. TERS, like SERS, is a method to enhance the Raman signal and, in principle, the mechanism of enhancement is the same. Scanning probe tips have dimensions of the order of tens of nanometers or less, and when metal coated, surface plasmon resonances can be optically induced, similar to the case for metallic nanoparticles. In TERS, the topography of the nanoscale environment of samples can be probed

by bringing the tip into close proximity with the area of the sample to be probed, but the Raman signal from the environment being probed by the tip is selectively enhanced by several orders of magnitude, swamping the spontaneous Raman from the remainder of the illuminated spot. Therefore TERS is a method which allows for very small areas or even individual molecules to be probed in a label free manner.

TERS has been used to investigate viral cell interaction[79], cytochrome-c states in isolated mitochondria[80], lipid and protein organisation in artificial cell membranes[81], as well as hemozoin crystal formation inside malaria infected red blood cells, as shown in Figure 2.2[82]. Figure 2.2. A –C show AFM images of infected red blood cells, highlighting the hemozoin crystals inside the cellular vacuole in C. Figure 2.2 D shows the TERS spectrum from the edge of the crystal deposits showing characteristic peaks associated hemozoin and the profile is compared to the SERS and RRS spectra of β-hematin in F and G . This study highlights TERS as a nanoscale technique with can be used to probe very specific areas which may have implications in disease. In this instance, TERS provides a potential method to study the interaction of quinoline anti-malarial drugs which are believed to preferentially bind to the edge of hemozoin crystals.

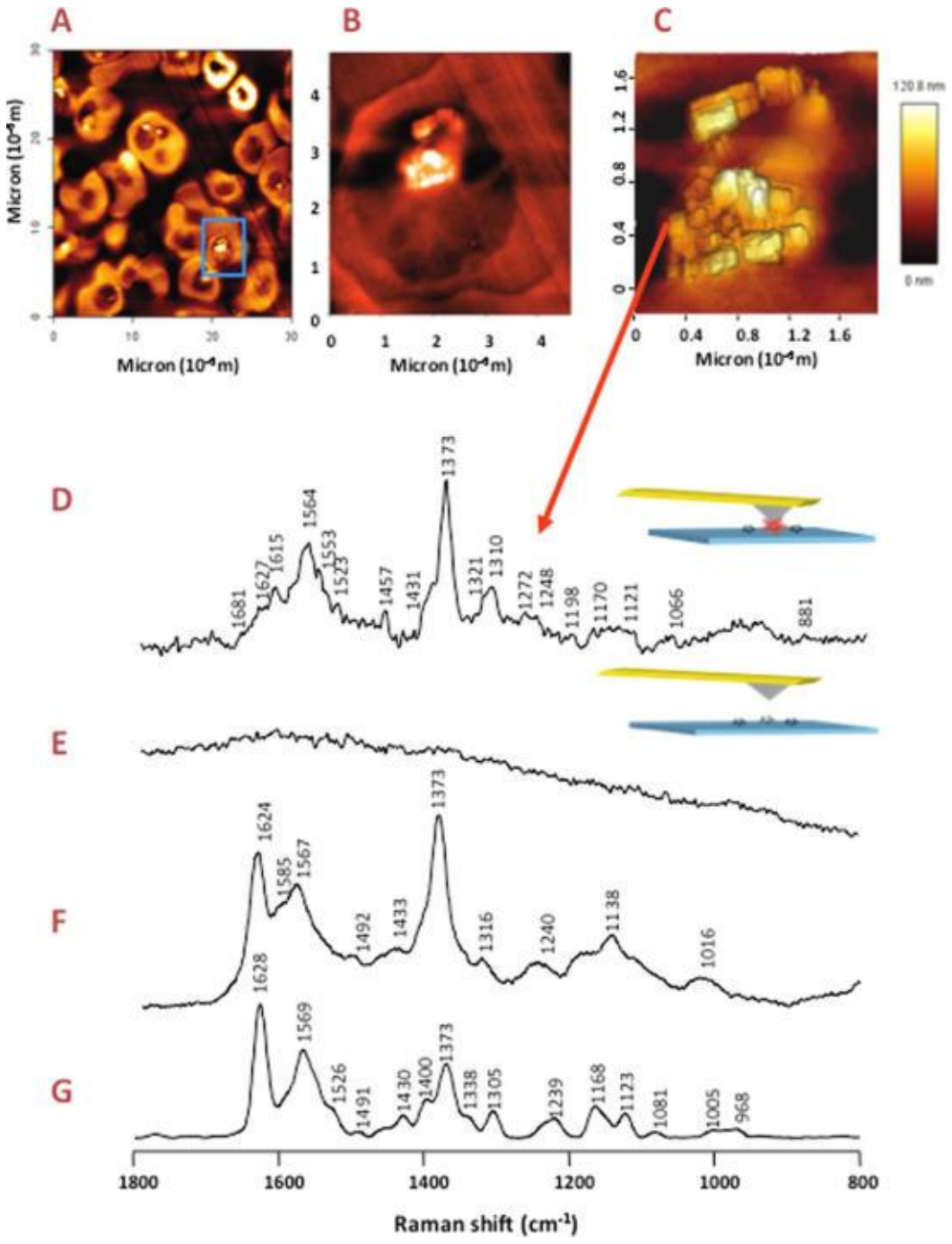


Figure 2.2 TERS probing hemozoin crystal formation inside malaria infected red blood cells. A – C show AFM images of infected red blood cells. D shows the TERS spectrum for the edge of the hemozoin crystal deposit, E is the spectrum of the tip following retraction from the cell, F SERS spectrum of β-hematin, G resonance Raman (RR) spectrum of β-hematin. Reproduced from [82].

TERS has also been used in the investigation of the interaction between cells and nanoparticles. Alexander and Schultz (2012) were able to show the interaction of individual antibody conjugated nanoparticles and cell surface bio molecules using TERS, with a similar sensitivity to SERS[83].

However, as TERS requires the use of AFM tips to enhance the signal, the method is restricted to being a surface classification technique and thus is of limited use for intracellular or indeed *ex-vivo* or *in-vivo* tissue analysis. While surface enhanced methods provide promise in a number of nanomedical areas, there are some caveats associated with these methods. Firstly, the probe must be capable of generating a surface enhancement of the Raman signal; this is only applicable to certain types of gold and silver particles or coated tips, as well as nanoaggregates of these metals. Additionally, these techniques require a considerable expertise in synthetic chemistry and design of probes or tips for specific target applications. Furthermore, reproducibility of the enhancement is also a concern, in particular with TERS, were the reproducibility of the tip characteristics is important in gathering reproducible spectra. Therefore it is important to consider that, while surface enhanced methods have been shown to be capable of nanoscale accuracy, these methods are heavily reliant on specifically designed nanoparticle sensors or probes and tips which in some way dilutes the label free aspect which spontaneous Raman spectroscopy provides.

2.5 Spontaneous Raman Spectroscopy

To differentiate it from the numerous variants of Raman spectroscopy which have emerged over the past decades, including SERS and TERS, the originally named phenomenon of Raman spectroscopy is now frequently called spontaneous

Raman spectroscopy. Spontaneous Raman spectroscopy has been used extensively over the past decades for a range of biomedical applications and is emerging as a viable alternative to gold standard protocols in medical diagnostics. Other uses include investigations in blood[84] and serum samples[85], investigations of human skin[38,39], cellular investigations[86–88], *in-vivo*[41] and *ex-vivo*[35] characterisations as well as studies of interaction of nanoparticles[36].

Importantly, these applications using Raman spectroscopy rely on the use of data analytical methods which aid in the classification and understanding of the data which has been acquired. This may entail the use of chemometric methods to cluster a data set so that one can see a cell or tissue as a distribution of similar spectra in a map. Multivariate statistical methods can be employed for the separation of two different classes of spectra e.g. a diseased and non-diseased state. A full description of such analytical methods is beyond the scope of this review. However, it is important to highlight how Raman spectroscopy and multivariate data mining approaches are commonly used together to investigate the biochemical nature of samples. Some good examples of where these statistical methods have been applied to Rama hyperspectral datasets can be found here[87,89,90].

Despite the extensive development of Raman spectroscopy for biomedical applications and the specific use of SERS using nanoprobes, not many studies have explored the use of spontaneous Raman spectroscopy for nanomedical applications. Of the reports that exist, some have aimed to look at probing cells for a toxic response[36,91], others have aimed to look at how nanomaterials behave in a cellular environment[34,92] and some have looked at degradation patterns of potential nanoparticle drug carriers[93].

The potential of Raman spectroscopy as a toxicological screening method has been demonstrated for the case of carbon nanotubes and their effects on human cells *in-vitro*. Kneif et al 2010, showed how the cellular spectral signatures differed between control and exposed cells due to changes in specific Raman spectral peaks of the cell nuclei. This method provided a way of investigating the toxic response of cells to nanomaterials in a truly label free manner, compared to more typical dye based cytotoxicity testing. In addition to detecting differences in response due to nanoparticle exposure, it was also possible to statistically compare the dose dependent responses of the Raman signatures with other gold standard toxicity tests, demonstrating the potential of the technique as a quantitative high throughput screening assay[36].

In a different type of study by Dorney et al 2012, the aim was to demonstrate the potential of Raman spectroscopy to visualise and investigate the interaction of polystyrene nanoparticles in cells. The purpose was to use these particles, which are often used as a standard in toxicity studies, as a model particle for further applications using Raman spectroscopy. In brief, the Raman spectroscopic signatures of the cells were mapped with a step size of 0.75μm over a region which contained both nuclear, perinuclear and cytoplasmic regions of the cell. Using a combination of K-means clustering and principal component analysis, it was possible to identify the localisation of the particles inside cells based on the intrinsic polystyrene signature and also to probe the chemical characteristics of the local subcellular environment[34]. A highlight of the results in shown in Figure 2.3 for cells incubated for 24hrs with polystyrene nanoparticles. The image in Figure 2.3 (i) shows the brightfield image (A) and the K-means image constructed for the Raman hyper spectral dataset (B). The polystyrene

nanoparticles are shown as green pixels in the image and the K-means average spectra are shown in Figure 2.3 (ii) A-D. The cluster associated with the green pixels clearly shows characteristic peaks associated with polystyrene when compared to a pure sample spectrum, Figure 2.3 (iii) A and B. The light blue and green clusters were then compared using Principal Component Analysis showing that the nanoparticles are located in lipid rich regions of the cell, which, by comparison with confocal fluorescence microscopy, was demonstrated to be the endoplasmic reticulum.

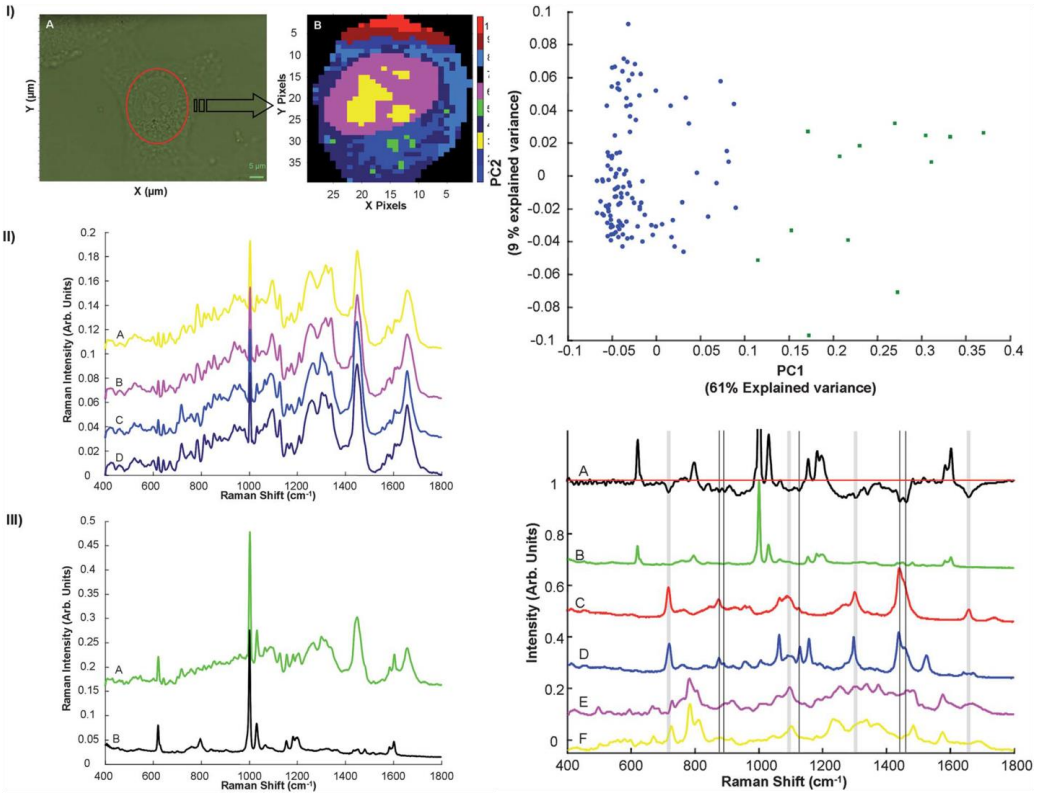


Figure 2.3. Identification of intracellular distributions of polystyrene nanoparticles using Raman spectroscopy. (i) A shows the brightfield and (B) K-means image of the cell. (ii) shows the K-means cluster average spectra associated with the clusters in the K-means image in the panel above, (iii) shows the K-means cluster spectrum associated with polystyrene nanoparticles (A) compared with a pure spectrum of polystyrene (B). The Right panels show a Principal Component Analysis scatter plot (top), differentiating the green (nanoparticle) and light blue clusters (cytoplasm), and the loading of Principal Component 1 (Bottom, A), suggesting the local environment surrounding the nanoparticles is lipid rich. Reproduced from [34].

Another recent study by Chernenko et al[94] aimed to investigate how different types of deuterated liposomal nanoparticles are distributed in cells. More specifically, it aimed to investigate how different chemical compositions affected how the liposomes associated with the mitochondrion. Notable in this study is the use of deuterated liposomes to enhance the ability to differentiate liposomes from endogenous lipids in the cell, based on the fact that the C-D vibrational frequency is significantly down shifted from that of the C-H stretch of

the very abundant intrinsic macromolecules of the cell. Another paper by the same group also looked at the degradation of polymeric nanoparticles over time in cells and concluded that poly lactic-*co*-glycolic acid (PLGA) and polycaprolactone (PCL) drug delivery systems are degraded and incorporated into the late endosomes of the Golgi system, based on spectral changes associated with the specific degradation patterns of the nanocarriers[93].

Spontaneous Raman spectroscopy has therefore already been demonstrated to be a chemically specific method for investigating nanoparticle interactions and also to probe the biochemical nature of cells. Notably, a number of biochemical features can be accessed simultaneously without the need for fluorescence or other labelling methods, or for costly cytotoxicological assays. It should be noted, however, that, based on current technologies, spontaneous Raman is a relatively weak effect, thus highlighting the attention which surface enhanced techniques such as SERS and TERS have received. Relatively weak signals can be compensated for by longer acquisition times, with maximum 2D scan times of the order of 40-80 mins for a 50µm*50µm area with a step size of 500nm for cellular data[95]. However, these scan speeds are largely dependent on the required signal to noise ratio and the step size used in image acquisition. For these reasons real-time imaging has not been realised to date.

Ultimately, for *in-vivo* applications, penetration depth is also an important consideration. In Raman microscopy, sensitivities are optimised by choice of objective, providing optical spatial resolution but limited penetration depth (~1-50µm). As Raman spectroscopy is an optical technique, the penetration depth is largely determined by the choice of wavelength of the source laser, and optimally this can be chosen in the near infrared region where tissue has a transmission

window. Absorption is largely governed by that of chromophores such as melanin (in skin) or haemoglobin across the visible, and by the overtones of OH vibrations in the near infra red regions. Scattering is an additional loss mechanism, but the development of Spatially Off-Set Raman Spectroscopy [77] using fibre probe rather than microscope objective delivery and collection optics, has exploited the fact that the signal from the deeper layers is scattered to a greater extent, to improve penetration depth sensitivities. CARS is another label free type of Raman scattering which can be used to probe bio and nanomedical scenarios and in recent years has seen a growth in applications in cells, tissues and *in-vivo* imaging. Using single wavelengths, imaging of large areas can be achieved at video rates.

2.6 CARS

Coherent anti-Stokes Raman spectroscopy (CARS) is a form of Raman spectroscopy whereby the anti-Stokes shifted Raman signal is used to probe the molecular bonds within a sample. The coherent process takes advantage of a third order non-linear optical phenomenon by which three beams are used to probe the sample. A fixed pump laser beam, a tunable probe beam are set at a frequency difference which is exactly equal to the frequency of a specific molecular vibration, resulting in the coherent build-up of a scattered signal on the anti-Stokes side of the pump laser frequency[96,97]. The signal can be orders of magnitude larger than a spontaneous Raman signal. Thus, CARS can be used to rapidly generate images of a particular biochemical distribution and therefore can be used in the generation of video rate image sequences of cells and tissues. To generate a full spectroscopic signature, however, the pump beam has to be tuned such that the difference frequencies scan the vibrational spectrum, a process

which can take considerable time, under current technological constraints. The nonlinear process is furthermore intensity dependent, requiring costly and notoriously temperamental short pulse lasers, whereas spontaneous Raman can be conducted with conventional steady state lasers.

In a biomedical context, the technique has been used to investigate a number of phenomena, also in conjunction with other methods such as immuno-fluorescent labelling. Primarily, CARS has been used in the study of the C-H stretch region which is most commonly associated with lipids in living organisms. Examples include the use of CARS for the study of atherosclerotic lesions[98], intracellular trafficking[99], drug delivery[100], cancer metastasis[101], quantitative imaging of lipid distributions in living *Caenorhabditis elegans*[102] as well as imaging of the axonal myelin both *in-vivo* and *ex-vivo*[103,104] CARS has also been used in the assessment of nanomaterials. Notably ,the technique has been used to study particle interaction in biological organisms, receptor mediated particle uptake[105] as well as the effects of particle size and coating on zebra fish embryos[106]. Moger *et al*[107] used CARS to investigate the interaction of metal oxide nanoparticles within the gills of rainbow trout, *Onchrhynchus mykiss*. They were able to show in a label free manner the translocation of TIO_2 particles across the epithelial membrane and into the capillaries in fish gill tissue. This is shown in figure 2.4, which illustrates the forward (a) and epi-CARS images (b) of exposed fish gills. The merged image shows the localisation of the particles in the gill tissues, revealing particle clumps in green.

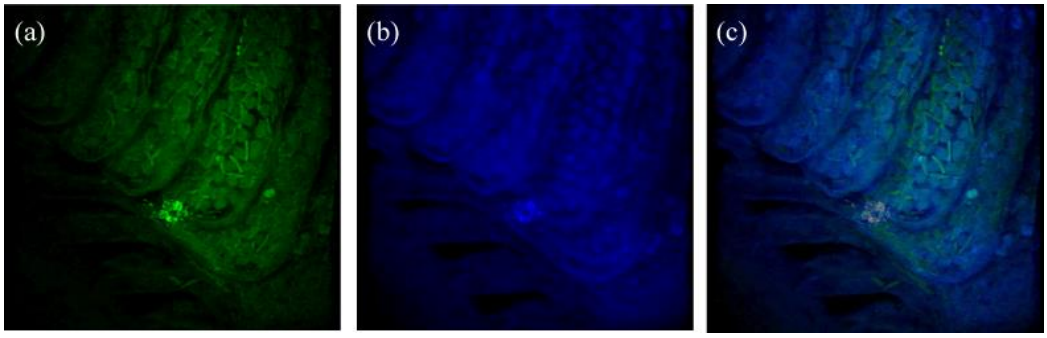


Figure 2.4 CARS images of the TiO_2 nanoparticle distribution in Onchrhynchus mykiss gills, (a) forward CARS image showing the nanoparticles, (b) epi-CARS image of the gill tissue and (c) merged forward and epi CARS image. Images reproduced from [107].

The method has also recently been used to investigate the mechanisms of oral uptake of Quaternary Ammonium Palmitoyl Glycol chitosan (GCPQ) nanoparticles. In this study, the particles were deuterated to shift the CH_2 stretching vibration located at 2840cm^{-1} to a CD_2 stretching vibration of 2100cm^{-1}. This allows for CARS to be carried out in the so called 'silent region' of the cell. Additionally second harmonic generation and two photon fluorescence were used to image the tissue containing nanoparticles. In doing this, Garrett et al. were able to examine chitosan uptake and recirculation in the gut by being able to target the nanoparticles with cellular precision to the gastrointestinal tract, liver and gall bladder, providing novel insights in the role of enterocytes and bile recirculation regarding chitosan nanoparticles[100,108]. Figure 2.5 shows the identification of the deuterated nanoparticles in green (2100cm^{-1}), which are highlighted by the arrows. Figure 2.5A and 2.5B show liver and stomach tissue respectively, with the C-D_2 resonance being used to identify the deuterated nanoparticles (2100cm^{-1}) in green and the C-H_2 (2845cm^{-1}) in red. Figure 2.5C shows a multimodal label free imaging approach combining CARS imaging (green), second harmonic generation (SHG) and two photon fluorescence (TPF)

in imaging nanoparticle interaction with jejunum tissue. Figure 2.5D and 2.5E show the use of a combination of CARS and TPF to image the ileum and duodenum respectively, while Figure 2.5F shows a combination of CARS, SHG and TPF of the gall bladder. These approaches show not only how CARS can be used to probe nanoparticle interactions, but also highlight how multiple imaging approaches can be combined in multimodal approaches to give different types of information building towards a more complete picture.

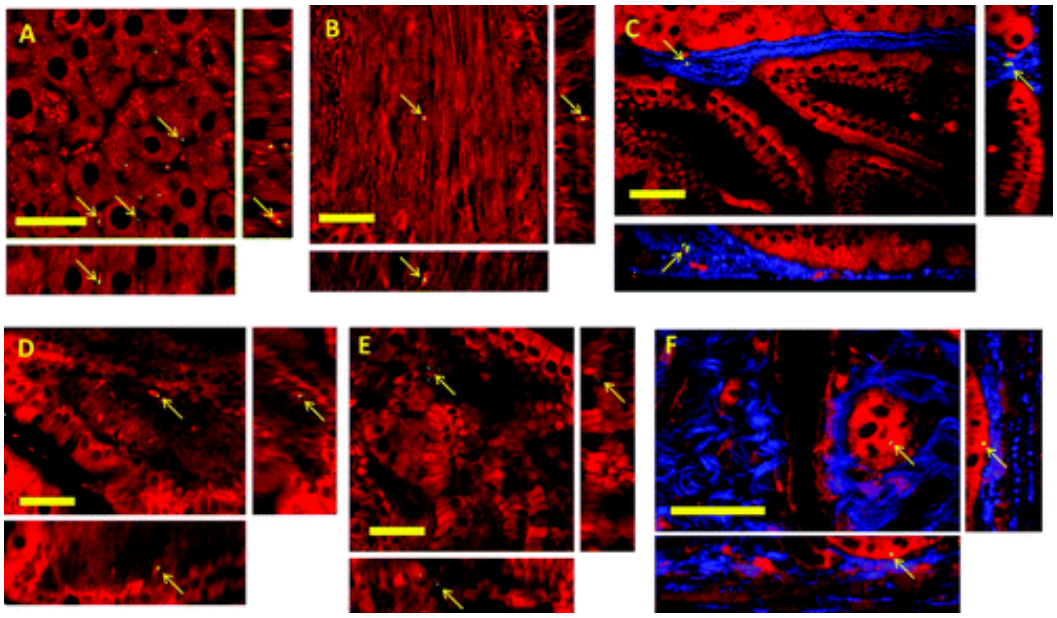


Figure 2.5 Epi-CARS images with contrast derived from CD_2 and CH_2 resonances in GCPQ nanoparticles at 2100 cm^{-1} (green) and 2845cm^{-1} (red) respectively. (A) Liver tissue. (B) Stomach tissue samples. (C) shows Jejunum tissue imaged with epi-CARS with contrast derived from the CD_2 resonance (green), SHG contrast derived from collagen (blue) and TPF contrast derived from endogenous fluorophores. (D) Ileum tissue imaged with epi-CARS with contrast derived from the CD_2 and TPF (red) (E) Duodenum imaged with epi-CARS with contrast derived from the CD_2 and TPF (red). (F) Gall bladder imaged with epi-CARS with contrast derived from the CD_2 resonance (green), SHG (blue) and TPF (red). Reproduced from[100]

Surface enhancement can also be exploited in the CARS format. Surface enhanced CARS (SECARS) has been used in conjunction with nanoparticles and has been shown to be capable of achieving greater signal enhancement than that of SERS or CARS alone. For biomedical applications, it has also been used for the detection of single molecules of deoxyadenosine and deoxyguanosine monophosphate (dAMP and dGMP)[109] and has also been used in immuno-histochemistry studies[110].

2.7 Conclusions and Outlook

This article has attempted to provide an overview of the current state of the art of the developing applications of Raman spectroscopic techniques in Nanomedicine. A recent review has dealt more broadly with the applications of these techniques in the investigation of the interaction of nanomaterials with complex biological systems[111]. The development of biomedical applications of vibrational spectroscopy, both Raman and IR, has been extremely active for the past two decades and more and the challenges to nanomedical applications are intrinsically linked, as indeed they are to those of the fundamental understanding of nanobio interactions in general.

As a molecular specific tool, Raman spectroscopy can potentially aid significantly to the understanding of nanobio interactions *in-vitro*. Even before interaction with the cell, it has been argued that the biological identity of the nanoparticle is determined by the surface coatings of the dispersion medium, the co-called protein corona[112]. While SERS active nanoparticles can be employed to probe this interaction acellularly, there is evidence that the nanoparticle medium interaction is very specific to the surface characteristics and size, and thus the use

of truly label free spontaneous Raman spectroscopy may lead to broader insights. In this context, the increased sensitivity of TERS may be of significant benefit. SERS has however demonstrated that the surface coating can evolve significantly after endocytosis of the nanoparticle [111], and this is a critical consideration in the bioavailability of surface functionalities, including release of active ingredients, which have been specifically designed for nanomedical applications.

As an confocal optical microscopic technique, Raman holds all the benefits of confocal fluorescence techniques, but has the potential advantage of being truly label free, adding the promise of reduced cost and sample processing requirements. SERS probes have demonstrated the potential to probe nanoparticle uptake, trafficking as well as the local environment, but these probes need to be specifically chemically tailored for the given application can so the technique cannot be considered to be truly label free. Spontaneous Raman spectroscopy is, on the other hand, an intrinsically weak phenomenon and cellular mapping is often a prolonged processes. Nevertheless, a number of cellular studies have been conducted which, although not specifically probing nanoparticles, may have implications in future nanomedical applications. For example, some studies have shown the application of Raman to drug delivery investigations[113,114] while other studies have identified sub cellular structures such as the mitochondrion as well as lipid rich regions which may be associated with the Golgi and endoplasmic reticulum[88]. Klein et al. used image registration and immuno-fluorescence to verify the locations of cellular organelles and also as a means of extracting the spectra which were specifically associated with the organelle[115]. These studies could be extended to look at nanoparticle trafficking studies, colocalizing the particle to an organelle in a label free manner, without using fluorescently labelled

nanoparticles or organelle stains. Although spontaneous Raman studies are commonly conducted on fixed cells, live cell spectral profiling has been demonstrated[40]. Image analysis is ultimately dependent on the reliability of multivariate chemometric techniques and simulated model systems can prove invaluable in validating their accuracy[83]. Increased acquisition rates can be achieved by systems custom designed for biological applications, and CARS potentially offers a route towards routine *in-vitro* screening of intracellular nanobio interactions, although its ability to rapidly screen the full spectrum is currently limited by the (tuneable) laser source technologies and applications are thus restricted by the need to identify specific spectral marker bands.

In terms of disease diagnostics, *ex-vivo* applications of Raman spectroscopy have received much attention. For the range of Raman modalities, however, mapping of large areas of tissue biopsies also suffer from issues of weak signals (spontaneous), specifically targeted probes (SERS), surface sensitivity (TERS) or the need for specific spectral markers (CARS). As a chemically specific probe, however, Raman techniques are particularly suitable for analysis of biomarkers of disease in biological fluids[7374] and this suitability is readily extended to applications in nanomedicine.

Raman scattering is fundamentally an optical technique and *in-vivo* applications are thus limited by the ability to access the area of interest. For dermal analysis, custom designed systems are commercially available which exploit the near infrared transmission window of skin, although, in a microscopic format, the penetration depth is further limited by the delivery optics, typically to some hundreds of microns. Advances in SORS have increased the depth resolution, and such technologies could prove invaluable tools for analysis of

transdermal nanodrug delivery or environmental exposure to nanoparticles. As an optical technique, Raman spectroscopy readily lends itself to endoscopic probes[116], however, and recent advances in such *in-vivo* probes may significantly impact on biomedical applications of Raman spectroscopy, including, inevitably Nanomedicine.

2.8 Future Perspectives

The field of nanotechnology is set to grow ever rapidly as new applications and avenues of research are explored over the coming decade. Crucially, characterisation and visualisation methods in a medical setting must develop in tandem, to access the applicability of such nanotechnology. Raman spectroscopy represents a method proven in the field of disease diagnostics and biomedical imaging and thus by extension holds the capability to progress the field of nanomedicine.

Spontaneous Raman spectroscopy provides a versatile and truly label free method which has seen success in a number of different medical applications, most notably in disease diagnostics. Key enabling technological developments in this context include endoscopic and other *in-vivo* probes. Relatively Low signal strengths currently limit the technique to small areas and/or long scan times, however, and continuing improvements in signal throughput and detector sensitivities are important. EU Directives limiting the use of animal models will put increasing emphasis on the development of in-vitro screening methods and Raman is a potential candidate for high content analysis of, for example, the efficacy and mode of action of novel chemotherapeutical agents of toxicants. The high optical resolutions obtainable make Raman particularly suitable for acellular

or subcellular studies of nanobio interactions. As the sensitivity of the Raman technique is intimately linked with the multivariate statistical data analysis methods, the quantitative specificities of these methods must be established. This can only be done if the true result is known, and in this context the use of specifically constructed model datasets may provide a quantifiable insight into how far Raman spectroscopy can be pushed in both a medical and nanomedical context.

SERS provides increased sensitivities to probe the nanoscale environment surrounding metallic nanoparticles. Although the technique is not truly label free, with the increased sensitivities achievable as well as the targeting potential of such probes, SERS may provide alternative imaging strategies for disease diagnostics *in-vivo,* as well as provide enhanced methods for the monitoring of human fluids such as serum and other metabolic excretions *ex-vivo*. SERS *in-vitro* may also prove a useful tool in probing the nature of the so called protein corona of nanoparticles in biological media and thus provide valuable insights into the surface behaviour of nanomaterials in a biological setting. Other enhancement methods such as TERS also provide novel insights into the nanoscale environment although they are limited by being mainly a molecular or surface specific technique.

Coupling these advances in spontaneous and surface enhanced Raman with the development of SORS and SESORS, some of the shortcomings in signal generation and depth penetration of Raman spectroscopy *in-vivo* may be overcome. In addition to the development of endoscopic and needle based probes which will increase access to the point of interest, realistic applicable *in-vivo* Raman studies in nanomedicine may not be too far away. CARS provides a

method which is capable of video rate scan speeds. However, as of yet the technique is not a spectroscopic imaging technique as it only allows for the probing of one particular wave number or vibrational marker at a time. The technique therefore requires a clearly identifiable biomarker for imaging, which may not be the case for all biomolecules. A CARS system that could provide a spectrum of the finger print region of the sample with similar real time imaging capabilities would be ideal. Specifically for CARS to progress as a spectroscopic imaging modality, advances in laser technology such as rapidly tunable lasers will need to develop in tandem. These advances would then open a myriad of applications for CARS imaging along the lines of spontaneous Raman imaging.

2.9 Executive Summary

Raman Spectroscopy: Raman spectroscopy is a well-established chemical analysis technique finding increasingly broader applications, particularly in biochemical analysis and disease diagnostics.

Surface/Tip enhanced Raman Spectroscopy: The techniques of SERS and TERS specifically probe the nanoscale and, although TERS is a topical/surface technique, SERS probes have already been used extensively for *in-vitro* and *in-vivo* studies. SERS probes are normally chemically functionalised according to the specific target, and so the technique is arguably not truly label free.

Spontaneous Raman Spectroscopy: As a truly label free technique, (spontaneous) Raman spectroscopy, coupled with multivariate analytical techniques potentially provides a probe of nanoparticles in cells/tissue, their nature of their local environment, and physiological changes. Unenhanced, the

signals are however relatively weak, and large scale mapping can be time consuming.

Coherent anti-Stokes Raman Spectroscopy: CARS is a nonlinear optical technique which is increasing in prominence for biomedical applications. Tuned to a specific vibrational frequency, it can scan large areas (cm^2) at video rates. Currently, however, it is not a spectroscopic technique and does not avail of the full biochemical information available, but relies on the presence of a specific spectral marker.

Outlook: The range of modalities of Raman spectroscopy potentially hold great promise for biomedical and nanomedical applications, although many technical challenges remain.

2.10 References

1 R. Freitas, *Nanomedicine, volume I: basic capabilities*, Landes Bioscience, Georgetown, TX, 1999.

2 A. Kumar, P. K. Vemula, P. M. Ajayan and G. John, *Nat. Mater.*, 2008, **7**, 236–41.

3 I. Perelshtein, G. Applerot, N. Perkas, J. Grinblat and A. Gedanken, *Chemistry*, 2012, **18**, 4575–82.

4 H. Yan, H. S. Choe, S. Nam, Y. Hu, S. Das, J. F. Klemic, J. C. Ellenbogen and C. M. Lieber, *Nature*, 2011, **470**, 240–4.

5 K. Nakano, K. Egashira, S. Masuda, K. Funakoshi, G. Zhao, S. Kimura, T. Matoba, K. Sueishi, Y. Endo, Y. Kawashima, K. Hara, H. Tsujimoto, R. Tominaga and K. Sunagawa, *JACC. Cardiovasc. Interv.*, 2009, **2**, 277–83.

6 O. Will, S. Purkayastha, C. Chan, T. Athanasiou, A. W. Darzi, W. Gedroyc and P. P. Tekkis, *Lancet Oncol.*, 2006, **7**, 52–60.

7 T. Skotland, T.-G. Iversen and K. Sandvig, *Nanomedicine*, 2010, **6**, 730–7.

8 N. Korin, M. Kanapathipillai, B. D. Matthews, M. Crescente, A. Brill, T. Mammoto, K. Ghosh, S. Jurek, S. a Bencherif, D. Bhatta, A. U. Coskun, C. L. Feldman, D. D. Wagner and D. E. Ingber, *Science*, 2012, **337**, 738–42.

9 V. M. Gaspar, I. J. Correia, A. Sousa, F. Silva, C. M. Paquete, J. a Queiroz and F. Sousa, *J. Control. Release*, 2011, **156**, 212–22.

10 H. Jin, J. F. Lovell, J. Chen, Q. Lin, L. Ding, K. K. Ng, R. K. Pandey, M.

Manoharan, Z. Zhang and G. Zheng, *Bioconjug. Chem.*, 2012, **23**, 33–41.

11 H. Bouwmeester, I. Lynch, H. J. P. Marvin, K. a Dawson, M. Berges, D. Braguer, H. J. Byrne, A. Casey, G. Chambers, M. J. D. Clift, G. Elia, T. F. Fernandes, L. B. Fjellsbø, P. Hatto, L. Juillerat, C. Klein, W. G. Kreyling, C. Nickel, M. Riediker and V. Stone, *Nanotoxicology*, 2011, **5**, 1–11.

12 B. G. Nair, T. Fukuda, T. Mizuki, T. Hanajiri and T. Maekawa, *Biochem. Biophys. Res. Commun.*, 2012, **421**, 763–7.

13 Q. Mu, N. S. Hondow, L. Krzemiński, A. P. Brown, L. J. Jeuken and M. N. Routledge, *Part. Fibre Toxicol.*, 2012, **9**, 29.

14 K. I. Willig, S. O. Rizzoli, V. Westphal, R. Jahn and S. W. Hell, *Nature*, 2006, **440**, 935–9.

15 M. J. Rust, M. Bates and X. Zhuang, *Nat. Methods*, 2006, **3**, 793–795.

16 X. Zhuang, *Nat. Photonics*, 2009, **3**, 365–367.

17 P. Sandin, L. W. Fitzpatrick, J. C. Simpson and K. a Dawson, *ACS Nano*, 2012, **6**, 1513–21.

18 F. Fazlollahi, S. Angelow, N. R. Yacobi, R. Marchelletta, A. S. L. Yu, S. F. Hamm-Alvarez, Z. Borok, K.-J. Kim and E. D. Crandall, *Nanomedicine*, 2011, **7**, 588–94.

19 E. Jan, S. J. Byrne, M. Cuddihy, A. M. Davies, Y. Volkov, Y. K. Gun'ko and N. a. Kotov, *ACS Nano*, 2008, **2**, 928–938.

20 J. Contreras, J. Xie, Y. J. Chen, H. Pei, G. Zhang, C. L. Fraser and S. F. Hamm-Alvarez, *ACS Nano*, 2010, **4**, 2735–2747.

21 J. Pawley, *Handbook of biological confocal microscopy*, Springer-Verlag,

Heidelberg, 3rd ed., 2006.

22 L. Aparicio-Ixta, G. Ramos-Ortiz, J. L. Pichardo-Molina, J. L. Maldonado, M. Rodríguez, V. M. Tellez-Lopez, D. Martinez-Fong, M. G. Zolotukhin, S. Fomine, M. a Meneses-Nava and O. Barbosa-García, *Nanoscale*, 2012, **4**, 7751–9.

23 Y. Yang, F. An, Z. Liu, X. Zhang, M. Zhou, W. Li, X. Hao, C. Lee and X. Zhang, *Biomaterials*, 2012, **33**, 7803–9.

24 http://www.invitrogen.com/site/us/en/home/support/Research-Tools/Cell-Staining-Tool.html.(date accessed 8/4/2019)

25 https://www.thermofisher.com/ie/en/home/brands/molecular-probes/key-molecular-probes-products/qdot.html (date accessed 8/4/2019)

26 J. P. Ryman-Rasmussen, J. E. Riviere and N. a Monteiro-Riviere, *J. Invest. Dermatol.*, 2007, **127**, 143–53.

27 D. J. Bharali, I. Klejbor, E. K. Stachowiak, P. Dutta, I. Roy, N. Kaur, E. J. Bergey, P. N. Prasad and M. K. Stachowiak, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 11539–44.

28 H. Cang, C. S. Xu, D. Montiel and H. Yang, *Opt. Lett.*, 2007, **32**, 2729–31.

29 A. Gajraj and R. Ofoli, *Langmuir*, 2000, 8085–8094.

30 K. Y. Win and S.-S. Feng, *Biomaterials*, 2006, **27**, 2285–91.

31 K. S. Suh H, Jeong B, Liu F, *Pharm Res.*, 1998, ;**15**, 1495–8.

32 A. Salvati, C. Aberg, T. dos Santos, J. Varela, P. Pinto, I. Lynch and K. a Dawson, *Nanomedicine*, 2011, **7**, 818–26.

33 G. Wang, A. S. Stender, W. Sun and N. Fang, *Analyst*, 2010, **135**, 215–21.

34 J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers and H. J. Byrne, *Analyst*, 2012, **137**, 1111–9.

35 F. M. Lyng, E. O. Faoláin, J. Conroy, a D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan and H. J. Byrne, *Exp. Mol. Pathol.*, 2007, **82**, 121–9.

36 P. Knief, C. Clarke, E. Herzog, M. Davoren, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2009, **134**, 1182–91.

37 A. Lattermann, C. Matthäus, N. Bergner, C. Beleites, B. F. Romeike, C. Krafft, B. R. Brehm and J. Popp, *J. Biophotonics*, 2013, **6**, 110–21.

38 S. Tfaili, C. Gobinet, G. Josse, J.-F. Angiboust, M. Manfait and O. Piot, *Analyst*, 2012, **137**, 3673–82.

39 Syed M. Ali ; Franck Bonnier ; Ali Tfayli ; Helen Lambkin ; Kathleen Flynn ; Vincent McDonagh ; Claragh Healy ; T. Clive Lee ; Fiona M. Lyng ; Hugh J. Byrne, *J. Biomed. Opt.*, 2013, **18**, doi:10.1117/1.JBO.18.6.061202.

40 F. Bonnier, a D. Meade, S. Merzha, P. Knief, K. Bhattacharya, F. M. Lyng and H. J. Byrne, *Analyst*, 2010, **135**, 1697–703.

41 H. Lui, J. Zhao, D. McLean and H. Zeng, *Cancer Res.*, 2012, **72**, 2491–500.

42 N. Colthup, L. Daly and S. Wiberley, *Introduction to infrared and Raman spectroscopy*, Academic Press, New York, 1975.

43 E. Smith and G. Dent, *Modern Raman spectroscopy: a practical approach*,

John Wiley and Sons, New York, 2005.

44 D.A. Long, *The Raman Effect: A Unified Treatment of the Theory of Raman Scattering by Molecules*, John Wiley and Sons, New York, 2002.

45 I. Lewis and H. Edwards, *Handbook of Raman spectroscopy: from the research laboratory to the process line*, CRC Press, 2001.

46 F. Parker, *Applications of infrared, Raman, and resonance Raman spectroscopy in biochemistry*, Springer-Verlag, Heidelberg, 1983.

47 L. A. Lyon, C. D. Keating, A. P. Fox, B. E. Baker, L. He, S. R. Nicewarner, S. P. Mulvaney and M. J. Natan, *Anal. Chem.*, 1998, **70**, 341–362.

48 W. M. Tolles, J. R. Mcdonald and A. B. Harvey, .

49 P. L. Stiles, J. a Dieringer, N. C. Shah and R. P. Duyne, *Annu. Rev. Anal. Chem. (Palo Alto. Calif).*, 2008, **1**, 601–26.

50 H. Byrne, G. Sockalingum and N. Stone, in *Biomedical Applications of Synchrotron Infrared Microspectroscopy*, Springer, 1st edn., 2011, pp. 105–142.

51 M. Fleischmann, P. Hendra and A. McQuillan, *Chem. Phys. Lett.*, 1974, **26**, 163–166.

52 D. Jeanmaire and R. Van Duyne, *J. Electroanal. Chem. …*, 1977, **84**, 1–20.

53 E. Blackie, E. Le Ru and P. Etchegoin, *J. Am. …*, 2009, **131**, 14466–72.

54 P. L. Stiles, J. a Dieringer, N. C. Shah and R. P. Van Duyne, *Annu. Rev. Anal. Chem. (Palo Alto. Calif).*, 2008, **1**, 601–26.

55 Y. Michael, H. Xu, S. Penn and R. Cromer, *Nanomedicine*, 2007, **2**, 725–734.

56 I. Chourpa, F. Lei and P. Dubois, *Chem. Soc. ...*, 2008, **37**, 993–1000.

57 W. Xie, P. Qiu and C. Mao, *J. Mater. Chem.*, 2011, **21**, 5190–5202.

58 Y. Chen, X. Zheng, G. Chen, C. He, W. Zhu, S. Feng, G. Xi, R. Chen, F. Lan and H. Zeng, *Int. J. Nanomedicine*, 2012, **7**, 73–82.

59 M. Schütz, D. Steinigeweg, M. Salehi, K. Kömpe and S. Schlücker, *Chem. Commun. (Camb).*, 2011, **47**, 4216–8.

60 X. Qian, X.-H. Peng, D. O. Ansari, Q. Yin-Goen, G. Z. Chen, D. M. Shin, L. Yang, A. N. Young, M. D. Wang and S. Nie, *Nat. Biotechnol.*, 2008, **26**, 83–90.

61 J. V Jokerst, A. J. Cole, D. Van de Sompel and S. S. Gambhir, *ACS Nano*, 2012, **6**, 10366–77.

62 R. Stevenson, A. Ingram, H. Leung, D. C. McMillan and D. Graham, *Analyst*, 2009, **134**, 842–4.

63 N. R. Isola, D. L. Stokes and T. Vo-Dinh, *Anal. Chem.*, 1998, **70**, 1352–6.

64 P. B. Monaghan, K. M. Mccarney, A. Ricketts, R. E. Littleford, F. Docherty, W. E. Smith, D. Graham and J. M. Cooper, *Anal. Chem.*, 2007, **79**, 2844–2849.

65 J. Kneipp, H. Kneipp, A. Rajadurai, R. W. Redmond and K. Kneipp, *J. Raman Spectrosc.*, 2009, **40**, 1–5.

66 J. Kneipp, H. Kneipp, W. L. Rice and K. Kneipp, *Anal. Chem.*, 2005, **77**, 2381–5.

67 J. Kneipp, H. Kneipp, M. McLaughlin, D. Brown and K. Kneipp, *Nano Lett.*, 2006, **6**, 2225–31.

68 J. Kneipp, H. Kneipp, M. McLaughlin, D. Brown and K. Kneipp, *Nano Lett.*, 2006, **6**, 2225–31.

69 J. Kneipp, H. Kneipp, B. Wittig and K. Kneipp, *J. Phys. Chem. C*, 2010, **114**, 7421–7426.

70 K. V. Kong, Z. Lam, W. D. Goh, W. K. Leong and M. Olivo, *Angew. Chem. Int. Ed. Engl.*, 2012, **51**, 9796–9.

71 J. Lin, R. Chen, S. Feng, J. Pan, Y. Li, G. Chen, M. Cheng, Z. Huang, Y. Yu and H. Zeng, *Nanomedicine*, 2011, **7**, 655–63.

72 R. McQueenie, R. Stevenson, R. Benson, N. MacRitchie, I. McInnes, P. Maffia, K. Faulds, D. Graham, J. Brewer and P. Garside, *Anal. Chem.*, 2012, **84**, 5968–75.

73 K. Kneipp and H. Kneipp, *Appl. Spectrosc.*, 2006, **60**, 322–334.

74 K. Ma, J. Yuen and N. Shah, *Anal. …*, 2011, **83**, 9146–9152.

75 N. Stone, K. Faulds, D. Graham and P. Matousek, *Anal. Chem.*, 2010, **82**, 3969–3973.

76 and A. W. P. P. Matousek, I. P. Clark, E. R. C. Draper, M. D. Morris, A. E. Goodship, N. Everall, M. Towrie, W. F. Finney, *Appl. Spectrosc.*, 2005, **59**, 393–400.

77 P. Matousek and N. Stone, *J. Biophotonics*, 2013, **6**, 7–19.

78 H. Xie, R. Stevenson, N. Stone, A. Hernandez-Santana, K. Faulds and D. Graham, *Angew. Chem. Int. Ed. Engl.*, 2012, **51**, 8509–11.

79 P. Hermann, A. Hermelink, V. Lausch, G. Holland, L. Möller, N. Bannert and D. Naumann, *Analyst*, 2011, **136**, 1148–52.

80 R. Böhme, M. Mkandawire, U. Krause-Buchholz, P. Rösch, G. Rödel, J. Popp and V. Deckert, *Chem. Commun. (Camb).*, 2011, **47**, 11453–5.

81 T. Deckert-Gaudig and V. Deckert, *Curr. Opin. Chem. Biol.*, 2011, **15**, 719–24.

82 B. Wood, E. Bailo, M. Khiavi and L. Tilley, *Nano Lett.*, 2011, **11**, 1868–1873.

83 K. D. Alexander and Z. D. Schultz, *Anal. Chem.*, 2012, **84**, 7408–14.

84 J. Shao, M. Lin, Y. Li, X. Li, J. Liu, J. Liang and H. Yao, *PLoS One*, 2012, **7**, e48127.

85 K. W. C. Poon, F. M. Lyng, P. Knief, O. Howe, A. D. Meade, J. F. Curtin, H. J. Byrne and J. Vaughan, *Analyst*, 2012, **137**, 1807–14.

86 F. Bonnier, P. Knief, B. Lim, a D. Meade, J. Dorney, K. Bhattacharya, F. M. Lyng and H. J. Byrne, *Analyst*, 2010, **135**, 3169–77.

87 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

88 C. Matthäus, T. Chernenko, J. a Newmark, C. M. Warner and M. Diem, *Biophys. J.*, 2007, **93**, 668–73.

89 F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 322–32.

90 M. Hedegaard, C. Matthäus, S. Hassing, C. Krafft, M. Diem and J. Popp, *Theor. Chem. Acc.*, 2011, **130**, 1249–1260.

91 P. Candeloro, L. Tirinato, N. Malara, A. Fregola, E. Casals, V. Puntes, G. Perozziello, F. Gentile, M. L. Coluccio, G. Das, C. Liberale, F. De Angelis and E. Di Fabrizio, *Analyst*, 2011, **136**, 4402–8.

92 H.-J. van Manen, Y. M. Kraan, D. Roos and C. Otto, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 10159–64.

93 T. Chernenko, C. Matthäus, L. Milane, L. Quintero, M. Amiji and M. Diem, *ACS Nano*, 2009, **3**, 3552–9.

94 T. Chernenko, R. R. Sawant, M. Miljkovic, L. Quintero, M. Diem and V. Torchilin, *Mol. Pharm.*, 2012, **9**, 930–6.

95 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

96 A. M. Zheltikov, *J. Raman Spectrosc.*, 2000, **667**, 653–667.

97 J. P. Pezacki, J. a Blake, D. C. Danielson, D. C. Kennedy, R. K. Lyn and R. Singaravelu, *Nat. Chem. Biol.*, 2011, **7**, 137–45.

98 H. Wang and I. Langohr, *Arter. Thromb Vasc Biol*, 2009, **29**, 1342–1348.

99 X. Nan, E. O. Potma and X. S. Xie, *Biophys. J.*, 2006, **91**, 728–35.

100 a Lalatsa, N. L. Garrett, T. Ferrarelli, J. Moger, a G. Schätzlein and I. F. Uchegbu, *Mol. Pharm.*, 2012, **9**, 1764–74.

101 T. T. Le, T. B. Huff and J.-X. Cheng, *BMC Cancer*, 2009, **9**, 42.

102 T. T. Le, H. M. Duren, M. N. Slipchenko, C.-D. Hu and J.-X. Cheng, *J. Lipid Res.*, 2010, **51**, 672–7.

103 H. Wang, Y. Fu, P. Zickmund, R. Shi and J.-X. Cheng, *Biophys. J.*, 2005,

89, 581–91.

104 Y. Fu, T. Huff, H. Wang, J. Cheng and H. Wang, *Opt. Express*, 2008, **16**, 19396–19409.

105 L. Tong, Y. Lu, R. J. Lee and J.-X. Cheng, *J. Phys. Chem. B*, 2007, **111**, 9980–5.

106 O. J. Osborne, B. D. Johnston, J. Moger, M. Balousha, J. R. Lead, T. Kudoh and C. R. Tyler, *Nanotoxicology*, 2012, 1–10.

107 J. Moger, B. D. Johnston and C. R. Tyler, *Opt. Express*, 2008, **16**, 3408–19.

108 N. L. Garrett, A. Lalatsa, I. Uchegbu, A. Schätzlein and J. Moger, *J. Biophotonics*, 2012, **5**, 458–68.

109 T.-W. Koo, S. Chan and A. a Berlin, *Opt. Lett.*, 2005, **30**, 1024–6.

110 S. Schlücker, M. Salehi, G. Bergner, M. Schütz, P. Ströbel, A. Marx, I. Petersen, B. Dietzek and J. Popp, *Anal. Chem.*, 2011, **83**, 7081–5.

111 D. Drescher and J. Kneipp, *Chem. Soc. Rev.*, 2012, **41**, 5780–99.

112 M. P. Monopoli, C. Aberg, A. Salvati and K. a Dawson, *Nat. Nanotechnol.*, 2012, **7**, 779–86.

113 H. Nawaz, F. Bonnier, P. Knief, O. Howe, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2010, **135**, 3070–6.

114 H. Nawaz, F. Bonnier, A. D. Meade, F. M. Lyng and H. J. Byrne, *Analyst*, 2011, **136**, 2450–63.

115 K. Klein, A. M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F.

Jamitzky, G. Morfill, R. W. Stark and J. Schlegel, *Biophys. J.*, 2012, **102**, 360–8.

116 M. Almond, J. Hutchings, C. Kendall, N. Stone, J. Day, N. Shepherd and H. Barr, *Gut*, 2011, **60**, A167–A168.

Chapter 3 Introduction to Raman spectroscopy and multivariate analytical methodologies applied to spectral datasets.

The following chapter contains sections from a journal article published in the Chemical Society Reviews, entitled: Spectral pre and post processing for Infrared and Raman spectroscopy of biological tissues and cells which Hugh J. Byrne was primary author, Peter Kneif was second and contributing author, Mark E. Keating was third contributing author with sections 3.2 – 3.9 used in the manuscript and Franck Bonnier was final contributing author.

Byrne HJ, Knief P, Keating ME, Bonnier F. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. Chem Soc Rev. DOI: 10.1039/c5cs00440c.

3.1 Introduction to Raman spectroscopy

Upon interaction with a material, light can be reflected, absorbed, or scattered. Raleigh scattering (elastic scattering) is when the scattered light is of the same frequency as the incident light. Raman scattering (inelastic scattering) is a result of light that is scattered by a material, whereby its frequency differs from that of the incident light, as a result of the interaction of a photon with the vibrations of a molecule.

In Raman scattering, the energy increase or decrease from the excitation is related to the vibrational energy spacing in the ground electronic state of the molecule, and therefore the Raman shift of the Stokes and anti-Stokes lines are a direct measure of the vibrational energies in a molecule. In Stokes Raman scattering, the molecule starts out in a lower vibrational energy state and, after the scattering process, ends up in a higher vibrational energy state. Therefore, the interaction of incident light with the molecule creates a vibration in a material and the scattered photon is reduced in energy.

In anti-Stokes Raman scattering, the molecule begins in a higher vibrational energy state and, after the scattering process, ends up in a lower vibrational energy state. Thus, a vibrational quantum in the material is annihilated as a result of the process and the scattered photon has an increased energy. The frequency differences between the Raman lines and the incident lines are characteristic of the scattering substance and are independent of the frequency of excitation.

The Raman effect arises from the coupling of the induced polarisation of scattering molecules, which is caused by the interaction of the electric field vector of the electromagnetic radiation with the molecular vibration modes. Light of frequency (ω_L) produces a polarisation in a material given by equation 3.1

$$P(\omega_L) = \chi(\omega_L) E_0 \cos \omega_L T \qquad \text{Equation 3.1}$$

where P is the polarisation, ω_L is the frequency of incident light, E is the electric field and $\chi(\omega_L)$ is the polarisability or susceptibility, normally considered a constant of the material associated with its electronic properties. However, at a finite temperature, a material is not at equilibrium and atoms will vibrate about there equilibrium position, R, along the normal coordinates with frequency ω_K, in accordance with a simple harmonic oscillator approximation. The displacement from equilibrium can be explained by equation 3.2

$$\Delta R(t) = \Delta R \cos(\omega_k t) \qquad \text{Equation 3.2}$$

The susceptibility to polarisation thus oscillates about its equilibrium value χ_0 and can be represented by equation 3.3

$$\chi_k(t) = \chi_0 + \Delta \chi_k \cos(\omega_k t) \qquad \text{Equation 3.3}$$

The polarisation now has the form as illustrated in equation 3.4

$$P(\omega_L, \omega_k) = \chi_0(\omega_L) E_0 \cos\omega_L t + \Delta \chi_k E_0 \cos(\omega_L) t \cos(\omega_k - \delta_k)$$

Equation 3.4

where δ_k takes into account any phase difference between the molecular vibration and the electric field oscillation. This may be written as equation 3.5

$$P(\omega_L, \omega_k) = \chi_0(\omega_L)E_0\cos\omega_L t + 1/2\Delta\chi_k E_0(\mathrm{COS}((\omega_L - \omega_k)t - \delta_k) + \cos((\omega_L + \omega_k)t + \delta_k))$$

Equation 3.5

Thus, the polarisation has the form

$$P = P(\omega_0) + P(\omega_0 - \omega_k) + P(\omega_0 + \omega_k)$$

Equation 3.6

An oscillating dipole will reradiate at the oscillation frequency, and thus the scattered light has three components. $P(\omega_0)$ gives rise to Raleigh scattering. $P(\omega_0 - \omega_k)$ corresponds to the subtraction of a vibrational quantum from the photon energy and the creation of a vibration and gives rise to the Stokes lines of the Raman spectrum. $P(\omega_0 + \omega_k)$ corresponds to the addition of a vibrational quantum to the photon by the annihilation of a vibration and results in the anti-Stokes lines of a Raman spectrum

3.2 Introduction to Multivariate Methods Applied to Raman Spectral Datasets.

Multivariate methods have become invaluable to a wide range of fields, including geology, pharmaceutical science, pharmacology, astrophysics, imaging, chemistry and the list goes on. Importantly, these methods allow for complicated and also in some instances very large datasets to be analysed and in effect they reduce the dimensionality and complexity of the data allowing for meaningful information to be extracted.

Specifically considering vibrational spectroscopic datasets, multivariate methods allow analysis of multiple spectra simultaneous and interdependently. This then allows for comparisons to be made between spectra and groups of

spectra within a dataset and to identify trends these may contain e.g. spectral markers of disease in control and non-control patients, identification of nanoparticle containing spectra, response to external agents such as drug or toxicants etc.

A Raman dataset usually consists of groups of spectra, which, depending on the sample and study being carried out, can be a set of random points, averaged spectra which can be the function of an external stimulus such as radiation, a chemical agent, nanoparticle etc. As the main focus of this work is centred on nanoparticle localisation and *in-vitro* drug screening using Raman spectral microscopy, multivariate statistical methodologies applied in this area will be discussed in more detail.

In Raman spectral microscopy, the dataset consists of a group of spectra which have been acquired via point mapping or raster scanning of a sample which may be cells or tissues, in *in-vitro* or *in-vivo* studies. As an imaging tool, much like fluorescent confocal laser scanning microscopy (CLSM), the sample has been scanned point by point, resulting in a dataset or in the case of spectral imaging a spectral hypercube. Unlike the simplicity of standard fluorescent imaging, whereby the dataset contains only one value per pixel, spectral hypercubes (as the name suggests) contain multiple data points per pixel which correspond to the spectrum acquired at that location. Similarly when spectra are acquired point by point each spectrum corresponds to the location sampled, without the spatial localisation achieved when imaging, with the benefit of the user knowing where the sample was acquired i.e. nucleus, cytoplasm etc.

As an imaging modality, to generate an image from this dataset, one must reduce the number of data points at each pixel to a single value. The simplest way

to achieve this is to form an image from one particular wavenumber in the spectrum. However, this method is somewhat flawed in that, if the peak in question corresponds to multiple biomolecules in a sample, it would be difficult to provide an accurate image of one particular composition. Nevertheless, in some instances, for example if a sample has a very distinctive peak, this may be the simplest way to generate an image. In a similar way, separation and classification can be achieved by single wavenumbers or ratios using point by point acquisition, although the multivariate nature of the technique is somewhat diluted if only a single wavenumber is used. Again if there is a prevalent change across the dataset this may be the simplest way to analyse the spectra.

As necessary, a myriad of methods have been developed to overcome this problem in spectral analysis. Using Raman spectroscopy as an example of where these methods are applied, clustering methods such as hierarchal clustering analysis (HCA), k-means clustering analysis (KMCA) and fuzzy c-means clustering (FCM) have been used to cluster spectra into groups and then based on these groups or classes, images and scatter plots can thus be generated following analysis. These clustering methods can be described as ‘hard clustering’ methods, in that each spectra is assigned a unique value and if a spectra has been assigned to one cluster it cannot be assigned to another.

Other methods have also been applied to Raman spectral analysis, including principal component analysis (PCA) and vertex component analysis (VCA). Both have been used for a number of applications. PCA has been used quite extensively to separate different sets of data based on the spectral variance present. This may be in a diagnostic setting and also in a spectral imaging sense.

VCA has also been used in this capacity although primarily in a Raman imaging setting.

Factor analysis methods have also been applied to Raman spectral datasets e.g. matrix factorisation (MF). In some instances, these methods are used to generate model spectra which in turn can be combined with other analytical approaches such as classical least squares analysis (CLSA) to evaluate, in a semi -quantitative way, the weighted contribution of each model spectrum to a particular spectrum for both images and individual groups of spectra.

This section of the introduction aims to give a brief description of some of the data mining approaches used to analyse vibrational spectroscopic data, focusing specifically on Raman spectroscopy, although examples from other spectral modalities such as IR spectroscopy will also be discussed in this context. The techniques which are explored more extensively in the thesis (e.g. KMCA, CLSA, PLSR and PCA) are described in more detail.

3.3 K-Means Cluster Analysis

K-means clustering analysis (KMCA) is a statistical method which aims to partition data into clusters based on similarity. K-means aims to minimise the sum of distances between spectral vectors S_j^i and cluster centroids m_k where J spectral vectors originally are randomly assigned to belong to a given cluster k with centroid m_k^1, see equation 3.7.

$$\sum(S_j^i - m_k) \quad \text{Equation 3.7}$$

Firstly, the method chooses a number of seed locations which serve as initial centroid locations in the dataset. Once a data point is assigned to one of the seed locations, it changes to a centroid which serves as a mean value of that cluster.

The assignment of data points to clusters is often based on the Euclidean distance between data point and centroid, although other methods of calculating the distance also exist[2]. After each spectrum has been assigned to a centroid, the distance is then recalculated between each point and centroid to see if any points are closer to another centroid location, whereupon, if the point is closer to another cluster centroid, then it is reassigned and both cluster centroids are changed as a result. This process is completed for all data points until there is no movement between clusters.

So, considering K-means from a Raman spectroscopic imaging perspective, an initial number of seed locations is chosen. The spectra are then assigned to one of the seed locations. Once all spectra have been assigned, the mean spectrum or centroid is calculated and the distance between each spectrum and centroid is calculated. The spectra are then reassigned if necessary and the process is iterated until no spectra change groups. Figure 3.1., shows a diagram highlighting the main steps in the K-means clustering algorithm.

In Raman spectroscopy, KMCA has seen a number of uses to separate spectra into clusters based on spectral similarities. As a Raman imaging tool, KMCA aims to separate each spectrum acquired in the image and assign it to a cluster. This assignment is termed ‘hard’ in that each spectrum is only assigned to one cluster. A good example of KMCA in Raman spectroscopy is shown in the work by Dorney et al[3]. Here, KMCA was used to identify regions in the Raman dataset which correspond to polystyrene nanoparticles, and differentiate them from neighbouring cytoplasm, as well as the nucleus and nucleolus. KMCA has also seen uses in other areas of Raman spectroscopy and spectral imaging such as the characterisation of skin layers. Good examples of KMCA as an spectroscopic

imaging reconstruction technique exist, the technique having been used in the investigation of a wide range of samples including tissue sections[4], cells[5,6] and in the analysis of human skin[7–9].

While this method has been shown to be useful in partitioning spectra into clusters, it is important to highlight that the method is not without certain caveats. Firstly, as the initial choice of centroid location can be subjective, the reproducibility of the method can in some instances be called into question i.e. if the initial starting point of the analysis changes then it is possible to end up with different results, so in practice if multiple datasets need to be compared all data should be analysed using the same centroid locations as these will change if analysis is carried out separately. Secondly, looking at the method to assess spectral imaging, each spectrum is assigned to only one cluster, and the cluster is represented by the average of all constituent spectra. As a Raman spectrum may represent a number of different biological entities in differing quantities, KMCA may be correct in grouping a spectrum based on lipidic distribution. However, it may misclassify a spectrum which also contains a small amount of another cluster's biochemistry. There is no weighting element introduced into the analysis, so one spectrum must belong to only one cluster even if multiple biochemical constituents are present. Thirdly, the number of clusters chosen is subjective and thus dependant on the user, if the incorrect number of clusters are chosen then spectra could be miss classified based on the loading from that group.

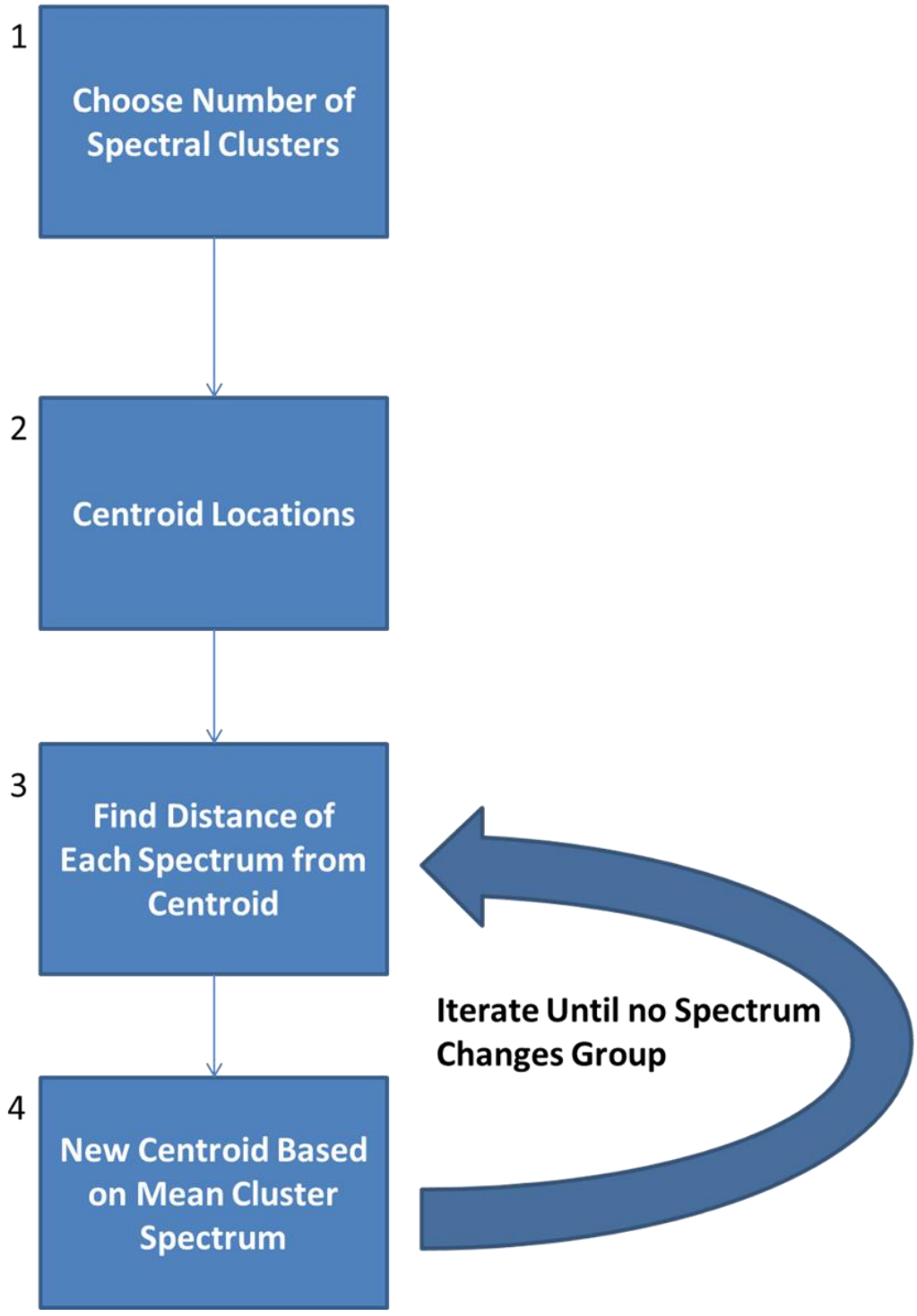


Figure 3.1 Schematic diagram outlining KMCA.

3.4 Fuzzy C – Means Clustering

Fuzzy C-means clustering is a method which is similar to KMCA in that it also assigns spectra to centroids in the datasets. However, unlike KMCA, the method is a soft clustering method, whereby each point or spectrum in the dataset is assigned a value from 0 to 1 for each particular cluster centre, with the value closest to 1 being representative of the cluster centre and 0 having no assignment. The algorithm developed by Bedzek et al[10], to calculate the degree of membership for each spectrum in the dataset results in a vector of the format $N_x N_y * C$ since each spectrum has C membership values. The co-efficients which describe membership to a particular cluster are defined by the following equation.

$$U_{iNS} = \frac{1}{\sum_{c=1}^{c} (d_{iNS}/d_{cNS})^{2/(m-1)}} \quad \text{Equation 3.8}$$

Where U_{iNS} is the membership of the sample N_s in one cluster, where d_{iNS} and d_{cNS} are the distances to the i^{th} and c^{th} cluster centres and m is the fuzziness factor between w and ∞.Therefore, by analysing the C centroid spectrum it is possible to extract chemical information which describes each reconstructed image. FCM has seen some usage in Raman spectroscopy although primarily as an imaging method [5,6].

3.5 Hierarchal Cluster Analysis

Hierarchal clustering analysis is another method which is commonly used for clustering spectral data and generating images. There are two main forms of HCA, agglomerative and divisive. Agglomerative HCA is the more commonly used method. Briefly, this method starts out with each data point or spectrum in a separate group or cluster. The method then aims to group each data point

together in an iterative process until there is only one cluster which contains all the data points. It is then possible to construct an image based on how these clusters are linked together. Often, the data can be represented using a two dimensional dendrogram which shows the linkage between each cluster. Divisive HCA on the other hand starts off with each spectrum in one cluster and then aims to separate each data point into one cluster. An example dendrogram is shown in Figure 3.2 An important point in relation to HCA is that, once a group of spectra has been assigned to a cluster or in the case of the agglomerative method merged into a cluster, the spectrum cannot be reassigned, unlike KMCA where the spectra can move clusters if closer to another centroid. This means that HCA results in a very definite grouping of spectra into clusters.

HCA is like KMCA in that the method is deemed to be a hard clustering method with each spectrum being assigned to a specific group. From an image reconstruction perspective and classification, this means that each pixel again can only be assigned to one specific biochemical grouping, which may not be reflective of the actual Raman dataset. HCA has been used in as a classification method in number of studies which include cellular studies[6] as well as in the investigation of vibrational spectroscopy in diagnostics[11].

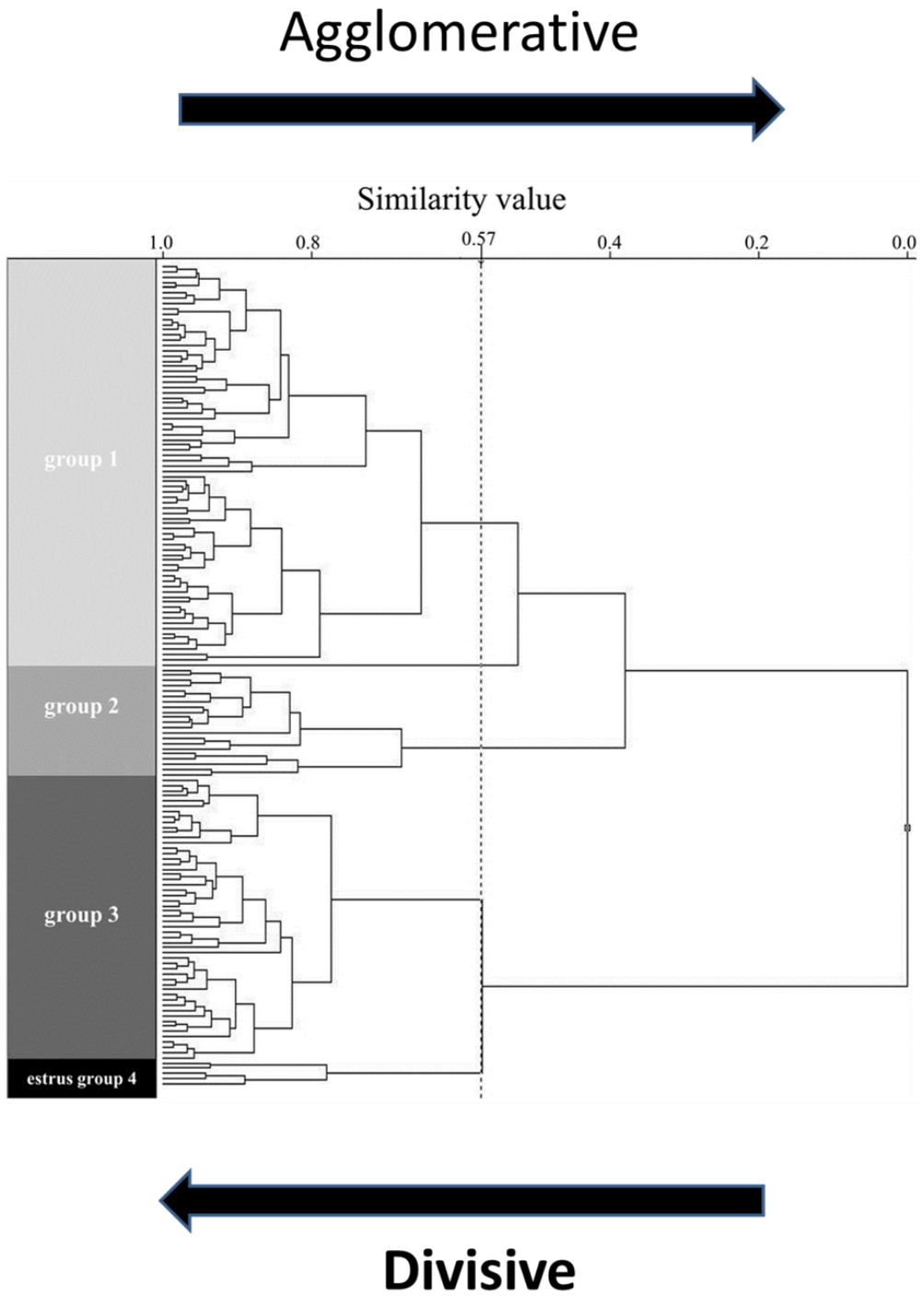


Figure. 3.2. Showing a HCA dendrogram and both divisive and agglomerative clustering. Adapted from[12].

3.6 Vertex Component Analysis

Vertex component analysis is another multivariate statistical method which is used in Raman spectral analysis[13]. The algorithm makes an assumption that, contained within the dataset, are pure endmember spectra which in turn can be used to describe all the other spectra in the dataset. From this, abundance plots can be generated via a linear combination of endmember spectra and constructed into images which are described by the biochemical information contained in these endmember spectra.

Assuming a linear mixing scenario each observed spectral vector is given by:

$$r = x + n = M\gamma\alpha + n$$

Equation 3.9

Where r is an L-vector (L is the number of bands), M = [m_1,m_2……m_p] is the mixing matrix (m_i denotes the ith endmember signature and p is the number of end-members present in the covered area), s = γα(γ is a scale factor modelling illumination variability due to surface topography), α = [α_1, α_2…… α_p]^T is the abundance vector containing the fractions of each end-member (.)^T stands for the vector transposed) and n model system additive noise.[13]

Recently, VCA has seen a number of applications in hyperspectral imaging using both IR and Raman spectroscopy, with applications including Raman histopathological imaging and also cellular studies including nano-bio interactions[4,14] Importantly, while this method can be used quite readily to reconstruct biochemical regions in the cell, like all methods it may be prone to error. Firstly, as highlighted in a paper by Chernenko et al.[14], endmember spectra

may contain mixtures of different biochemical components and while this may be reflective of the actual nature of the sample, may lead to inaccuracies in interpretation. Additionally, the method makes a large assumption that the most extreme spectra in the dataset are the most reflective of pure component spectra, which may not be the case in complex biological spectra.

3.7 Principal Component Analysis

PCA is a method which aims to reduce the dimensionality of the data to describe the variation present in a dataset, whereby the first principal component is a description of the maximum variance present in the dataset, the second describes the second most variance…etc. The principal component scores can then be described by the loading vector which is an explanation of this variance. In a Raman spectroscopy context the scores represent values which correspond to a loading spectrum which contains peaks, both positive and negative which explains the spectral variation in the dataset.

This tool can be quite useful for providing a method to separate spectra into groups e.g. diseased and non-diseased[15]. It has also been used to reconstruct images[6,16], i.e. a variance plot based on the loadings plot. However, as these loadings plots may often contain a number of spectral features corresponding to different cellular biochemistry, interpretation can be difficult and it is quite possible to misinterpret. Bonnier et al have shown that pairwise PCA of clusters identified by KMCA can provide a clearer picture of the specific biochemical differences between region[17].

In this thesis Seeded Principal Component Analysis (SePCA) is introduced as a novel multivariate analysis variant to address some of the

limitations of the application of PCA in bio-spectroscopy in systematically varying datasets. Using simulated data based on experimental spectra of *in-vitro* exposure to varying doses of the chemotherapeutic agent, cis-platin, standard and SePCA are compared, firstly based on their ability to differentiate the responses to different exposure doses, and secondly to assess the accuracy of the loadings that are used to describe the systematic variations of biochemistry underlying the differentiation. Further insights are also garnered on the use of 1st and 2nd derivative spectra and the impact this mathematical transformation has on the ability of the algorithm to separate and describe the spectral origin of differentiation of spectral datasets. The implications of this novel variant of PCA are discussed in the context of screening for drug efficacy *in-vitro* as well as biomedical classification for disease diagnostics.

3.8 Partial Least Squares Regression

Partial least squares regression (PLSR) is an analytical technique which aims to match a test data set to a series of targets. In brief, the method aims to create a model dataset which relates a spectral dataset to a series of test points or targets (i.e. concentration, dosed). The spectral data (X matrix) is thus related to the targets (Y matrix) according to the linear equation;

$$Y = XB + E$$

Equation 3.10

where B is a matrix of regression coefficients and E is a matrix of residuals[18]. A good practical example of this method in action in Raman spectral data is outlined in two studies by Nawaz et al.[16,17], in which the aim was to investigate the capability of Raman as a technique to study drug interactions in cells and the

physiological response. Looking specifically at cis-platin as an example chemotherapeutic drug, these studies were able to extract information relating to drug action in the cells via regression of the Raman dataset against cytotoxicological data and dose. Features were extracted from the Raman spectra which correspond to changes to protein conformation and structural alterations of DNA[19,20].

Importantly, while these studies show the potential of Raman spectroscopy and PLSR as tools for studying drug interaction, PLSR used in this capacity is only relevant if the processes studied are in themselves linear. However, most pharmacological actions are non-linear processes and thus using a linear method to model a non-linear process may be subject to error. Thus, additional forms of validation of these methods in a spectral setting are necessary.

In this thesis, PLSR is investigated using simulated datasets based on previously published data. In this way the application of the PLSR algorithm is investigated and the limits and sensitivities are explored using a simulated dose and cytotoxicological target dataset, providing a methodology for the assessment of multivariate approaches used for Raman *in-vitro* screening.

3.9 SVM

Support vector machines is a classifier which aims to partition data to give a separation between control and sample. Generally, the algorithm is used in conjunction with PCA, whereby the coefficient values are used to build a model which is then used to classify samples. Initially data with a known classification is used to train the model i.e. control vs. cancer. Once the model has been trained, samples with an unknown grouping are classified based on their affinity for each

group. In this way, it is possible to classify samples as one group or the other when the grouping is unknown.

3.10 Concluding Remark

While the list of methods is far from exhaustive, it highlights some of the commonly used analytical methods in Raman spectral analysis. Some of these methods have certain caveats associated with them and thus may be prone to error for certain applications. In the following chapters, some of the possible issues associated with these methods in an *in-vitro* Raman setting are explored, primarily using simulated datasets which are based on real experimental data. Novel variants and methods are also explored to tackle some of the issues which have arisen while investigating these methods, with the central thesis focusing on the use of simulated datasets in assessing the validity, accuracy and applicability of multivariate statistical methodologies for *in-vitro* Raman screening and beyond. ‘

In the following chapters, some of the possible issues associated with these methods in an *in-vitro* Raman setting are explored, primarily using simulated datasets which are based on real experimental data. More detailed descriptions of the underlying theories of PLSR analysis (Chapter 4), PCA (Chapter 5) and CLS analysis (Chapter 6) are provided. Novel variants and methods are also explored to tackle some of the issues which have arisen while investigating these methods, with the central thesis focusing on the use of simulated datasets in assessing the validity, accuracy and applicability of multivariate statistical methodologies for *in-vitro* Raman screening and beyond.

3.11 References

1 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

2 A. Rencher and W. Christensen, *Methods of multivariate analysis*, John Wiley and Sons, New York, 3rd edn., 2012.

3 J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers and H. J. Byrne, *Analyst*, 2012, **137**, 1111–9.

4 N. Bergner, B. F. M. Romeike, R. Reichart, R. Kalff, C. Krafft and J. Popp, *Analyst*, 2013, **138**, 3983–90.

5 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

6 M. Hedegaard, C. Matthäus, S. Hassing, C. Krafft, M. Diem and J. Popp, *Theor. Chem. Acc.*, 2011, **130**, 1249–1260.

7 Syed M. Ali ; Franck Bonnier ; Ali Tfayli ; Helen Lambkin ; Kathleen Flynn ; Vincent McDonagh ; Claragh Healy ; T. Clive Lee ; Fiona M. Lyng ; Hugh J. Byrne, *J. Biomed. Opt.*, 2013, **18**, doi:10.1117/1.JBO.18.6.061202.

8 S. M. Ali, F. Bonnier, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T. C. Lee, F. M. Lyng and H. J. Byrne, *Anal. Methods*, 2013, **5**, 2281.

9 S. M. Ali, F. Bonnier, A. Tfayli, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T. Clive Lee, F. M. Lyng and H. J. Byrne, *J. Biomed. Opt.*, 2013, **18**, 061202.

10 J. C. Bezdek, 1984, **10**, 191–203.

11 P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J. H. Shanks, M. Brown, N. W. Clarke and P. Gardner, *Analyst*, 2012, **137**, 1370–7.

12 K. Kinoshita, M. Miyazaki, H. Morita, M. Vassileva, C. Tang, D. Li, O. Ishikawa, H. Kusunoki and R. Tsenkova, *Sci. Rep.*, 2012, 2, 856.

13 J. M. P. Nascimento, S. Member and J. M. B. Dias, *IEEE Trans. Geosci. Remote Sens.*, 2005, **43**, 898–910.

14 T. Chernenko, R. R. Sawant, M. Miljkovic, L. Quintero, M. Diem and V. Torchilin, *Mol. Pharm.*, 2012, **9**, 930–6.

15 F. M. Lyng, E. O. Faoláin, J. Conroy, a D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan and H. J. Byrne, *Exp. Mol. Pathol.*, 2007, **82**, 121–9.

16 M. MiljkoviÄ‡, T. Chernenko and M. Romeo, *Analyst*, 2010, 2002–2013.

17 F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 322–32.

18 P. Knief, C. Clarke, E. Herzog, M. Davoren, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2009, **134**, 1182–91.

19 H. Nawaz, F. Bonnier, P. Knief, O. Howe, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2010, **135**, 3070–6.

20 H. Nawaz, F. Bonnier, A. D. Meade, F. M. Lyng and H. J. Byrne, *Analyst*, 2011, **136**, 2450–63.

Chapter 4 Multivariate statistical methodologies applied in biomedical Raman spectroscopy: Assessing the validity of partial least squares regression using simulated model datasets.

The following chapter contains sections from a journal article submitted to Analyst, entitled: Multivariate statistical methodologies applied in biomedical Raman spectroscopy: Assessing the validity of partial least squares regression using simulated model datasets which Mark E. Keating was primary author and responsible for data analysis, writing and formatting of the paper, Haq Nawaz was second author and contributed code and spectral data, Franck Bonnier was third author and co-supervisor and Hugh J. Byrne was final author and supervisor.

Keating ME, Nawaz H, Bonnier F, Byrne HJ. Multivariate statistical methodologies applied in biomedical Raman spectroscopy: assessing the validity of partial least squares regression using simulated model datasets. Analyst. 2015 Mar;16;140(7):2482-92. doi: 10.1039/c4an02167c.

4.1 Abstract

Raman spectroscopy is fast becoming a valuable analytical tool in a number of biomedical scenarios, most notably disease diagnostics. Importantly, the technique has also shown increasing promise in the assessment of drug interactions on a cellular and subcellular level, particularly when coupled with multivariate statistical analysis. However, an important consideration, both with Raman spectroscopy and the associated statistical methodologies, is the accuracy of these techniques and more specifically the sensitivities which can be achieved and ultimately the limits of detection of the various methods. The purpose of this study is thus the construction of a model simulated data set with the aim of testing the accuracy and sensitivity of the partial least squares regression (PLSR) approach to spectral analysis. The basis of the dataset is the experimental spectral profiles of a previously reported Raman spectroscopic analysis of the interaction of the cancer chemotherapeutic agent cis-platin in an adenocarcinomic human alveolar basal epithelial cell- line, *in-vitro*, and is thus reflective of actual experimental data. The simulated spectroscopic data is constructed by adding known perturbations which are independently linear in drug dose, as well as cytological response, experimentally determined by the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) cytotoxicity assay. It is demonstrated that, through appropriate choice of dose range, PLSR against the respective targets can differentiate between the spectroscopic signatures of the direct chemical effect of the drug dose and the indirect cytological effect it produces.

4.2 Introduction

Over the past couple of decades, vibrational spectroscopy (in particular Raman and infrared absorption) has emerged as a powerful tool for biomedical applications. The numerous studies explore applications such as disease diagnostics[3–6], cellular imaging[7–10], the study of drug[1,2,11]and nanoparticle interactions[12–14] on a cellular and sub-cellular level, to name but a few. In both modalities, the spectrum of tissue or cells contains a wealth of information, representing as it does the combined molecular fingerprints of the ensemble of biomolecules contained in the sample, and only in the simplest of cases can a valid interpretation be made by visual inspection of the spectrum. Multivariate statistical methods are thus critical in the analysis, interpretation and representation of the complex information contained within. However, given the critical nature of the outcomes of the application, whether in terms of medical diagnostics or in preliminary screening of drug efficacy and action mechanisms, it is imperative that the combination of spectroscopic techniques and multivariate analysis are rigorously and quantifiably validated. Such validation can also establish realistic limits to what is often purported as a high content screening methodology. To this aim, the use of simulated datasets based on experimental studies can play a crucial role[14,15].

A multitude of multivariate analytical methods exists, each of which aims to simplify complex bio-spectroscopic information and provide a tool with which to draw conclusions about the state of the sample. These include Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Vertex Component Analysis (VCA), Spectral Cross Correlation Analysis (SCCA), K-

means Clustering Analysis (KMCA), Hierarchal Cluster Analysis (HCA) to name but a few. Importantly, there also exists a number of variants of these methods which differ slightly and can give, in some instances, different answers[14,16,17].

Recently, regression modelling (e.g. Partial Least Squares Regression, PLSR) has seen a number of biomedical uses in both Raman and IR spectroscopies. The core idea of using this method is to investigate the spectral variability as a function of a systematic conditional change such as radiation dose[18] or viral infection[19]. PLSR can be employed to construct predictive models for spectral response as a function of the target variable. Therefore, an unknown dose or degree of infection can be determined from its spectrum, having obvious potential clinical applications. Furthermore, feature selection techniques such as PLSR co-efficient, Jack-Knifing (JK) and genetic algorithms, amongst others[20], can be employed to identify the most statistically relevant spectral changes, such that the biological mechanisms underlying the spectral changes can explored and understood. Importantly, there are many variants of the PLSR algorithm and, in some instances; hybrid methods which use a combination of two statistical tools in order to extract relevant chemical information have been employed. Although these methods have been applied to a wide range of studies, the details are beyond the scope of this paper although good examples can be found in literature[1,2,18,21–24]

The potential of Raman spectroscopic microscopy for initial screening of chemotherapeutic efficacy and mechanism of action has been demonstrated by Nawaz *et al.*[1,2,23]. Taking the interaction of cis-platin with the human lung adenocarcinoma cell line, A549, *in-vitro*, as an example, PLSR of Raman spectroscopic datasets was reported to identify and differentiate the direct effects

of cis-platin on the cellular biochemistry as a function of drug concentration (dose) and the resultant toxicological response as measured by the MTT cytotoxicity assay. This simultaneously provides a parallel gold standard technique to compare to the spectroscopic endpoint as well as range finding for the initial dose response curve i.e. establishing values of Inhibitory Concentrations (IC) etc. In an operational model of pharmacological agonism, the former is a linear process, whereas the latter results in the more complex sigmoidal response of cell populations to drug exposure[25]. PLSR against the drug concentration returned changes in the Raman peaks associated with both conformational and chemical changes in DNA, while changes to the lipid and protein distributions were dominant when the data was regressed against the cytotoxicological end point, indicating the biochemical changes associated with the resultant cytological response to the interaction with cis-platin. The statistic relevance of the results were confirmed using the JK approach.

The potential to differentiate the direct chemical effects from the subsequent cytological responses opens the way to the use of the techniques to visualise and interpret the mode of action of chemotherapeutic agents intracellular and to quantify the efficacy to produce the desired cellular response in a single truly label free measurement. The emergence of ever higher throughput spectrometers would enable real-time and time resolved visualisation of the respective processes as they evolve. Notably, however, while the studies of Nawaz *et al*. show great promise towards this end, the technique is as yet un-validated. The expected changes in the spectra with concentration and toxicological endpoint are inferred, based on prior knowledge about the biological action of cis-platin in the model *in-vitro* system. This leads to a difficulty when

trying to confirm the validity of the method or compare two different methods to quantitatively assess the sensitivity, accuracy and specificity of the technique.

Here, we aim to validate the application of these methodologies using simulated datasets based on the previously published experimental results of Nawaz *et al.*. In particular, we aim to test the ability of PLSR to model and thus extract spectroscopic variations (based on the regression co-efficient) which vary systematically as a function of different targets. Thus, the study will confirm whether the method is capable of extracting and differentiating spectroscopic features which differ based on linear or non-linear changes of the targets. Additionally, the accuracy or fidelity of the method in extracting systematically varied features will be explored as the spectral perturbations introduced decrease in magnitude, exploring the sensitivity of the method. Thus, the overarching aim is to establish the validity of the algorithms applied to Raman spectral datasets containing changes pertaining to the direct and indirect effects of the anti-cancer drug cis-platin *in-vitro*. For the purposes of this study, we propose the use of a modelled simulated dataset. The dataset is constructed based on experimental observations, but the systematic spectral variation that is introduced is known precisely and thus an exact and complete assessment of the method can be carried out.

4.3 Methods

4.3.1 Experimental

Experimental results were obtained as described in previous publications by Nawaz *et al*[1,2] which investigated Raman spectroscopy as a tool to study cis-

platin-cellular interactions *in-vitro*. The experimental methods are described in detail in the publications, but are summarised in brief as follows.

Human lung adenocarcinoma (A549) cells were routinely cultured at 37 °C, 5 % CO_2 in DMEM F12 supplemented with 10% FBS, 1% pen/strep and 2mM l-glutamine. Cells were cultured until 70-80% confluence and plated on quartz substrates for Raman spectroscopy. A standard MTT assay, using a concentration range of 0.05µM – 50µM, was used to assess the toxicity of cis-platin to provide a comparison to Raman spectroscopy. This was carried out in standard 96 well plates and experiments were all completed in triplicate. This range resulted in a sigmoidal variation in cell culture viability over the range ~90% to ~20%, from which the Inhibitory Concentration (IC_{50}) of cis-platin in A549 cells *in-vitro* was determined to be 1.2 ± 0.2 µM.

Cis-platin, at varying concentrations in the range 0.05 µM - 50µM, was added to cells and Raman microscopic measurements of cells exposed to each dose, including unexposed control, were acquired at a source wavelength of 785nm for both nuclear[1]and cytoplasmic regions[2] . The PLSR approach was used to model the spectroscopic data as well as to select and distinguish the relevant features indicative of the chemical effects of cis-platin and the cellular response to cis-platin via a regression against dose and the MTT cytotoxicity endpoint respectively. By examination of the regression co-efficient, it was possible to discern the major features responsible for model construction.

In this work, these experimental spectral datasets are employed to construct semi-realistic simulated data to probe the reliability, sensitivity and quantitative nature of these methods when applied to drug-interaction studies. More details of the experimental set up can be found in Nawaz *et al.* [1,11]

4.3.2 Partial Least Squares Regression

PLSR is a multivariate statistical method which aims to establish a model that relates the variations of the spectral data to a series of relevant targets. The spectral data (X matrix) is thus related to the targets (Y matrix) according to the linear equation Y = XB +E, where B is a matrix of regression coefficients and E is a matrix of residuals. The PLSR algorithms used in this study have been previously published elsewhere[1,2,18,22] and are based on scripts written in house using Matlab 7.2 (The Mathworks Inc.). The algorithm allows for the construction of a regression model which can be used to predict the outcome in a number of different situations. In this case, the examples used are concentration and MTT response, and therefore the algorithm can be used to predict for example the toxicological response of a particular drug dose.

Latent variables (LV's) in PLSR modelling are a series of underlying variables which aim to describe the behaviour of the modelled system. The exact number of latent variables which are necessary to build an entirely accurate model is not known *a priori*. However, it is one of the goals of PLSR models to accurately predict the number necessary to build a robust and accurate model[26]. Predicting the number of LVs which will build an accurate model is usually achieved during the cross validation step, typically using the root mean squared error of cross validation (RMSECV) as a metric for latent variable selection.

4.3.3 Spectral Constructs

Spectral constructs were generated for the purpose of imparting a known perturbation to the dataset which could be systematically varied to evaluate the capability of the PLSR modelling to accurately predict and extract spectral variations correlated to a known external variable, in this case, drug dose and the

resultant cytological changes. Using the original datasets of Nawaz *et al.*, derived from the nuclear and cytoplasmic regions, specific spectral changes were identified in the mean difference spectra of a 3μM exposed cell population versus the unexposed control (Figure 3, of reference 10, Figure 4 of reference 11). In this way, spectral constructs were generated from the changes in the spectra of the nuclear region, including increases in the characteristic A form of DNA peak at 807 cm^{-1} and the B form peak at 833 cm^{-1} and a change in the C-H deformation at 1449 cm^{-1} (Figure 4.1A) and in the cytoplasmic region, containing the following peak changes or shifts; a change in the amide 1 band at ~1661 cm^{-1}, a decrease in the C-C stretch intensity at ~939 cm^{-1} and an increase in the tryptophan peak at 731 cm^{-1} (Figure 4.1B). The relative intensities of the peaks in each construct were derived from the experimental difference spectra at a cis-platin exposure dose of 3μM[10] and were normalised for concentration (Figure 4.1A) and a loss of viability at that concentration of 0.52 [10] (Figure 4.1B). Different weightings of these spectral constructs (termed hereafter the Concentration and Viability construct respectively) were then added to a control dataset as described in the following section.

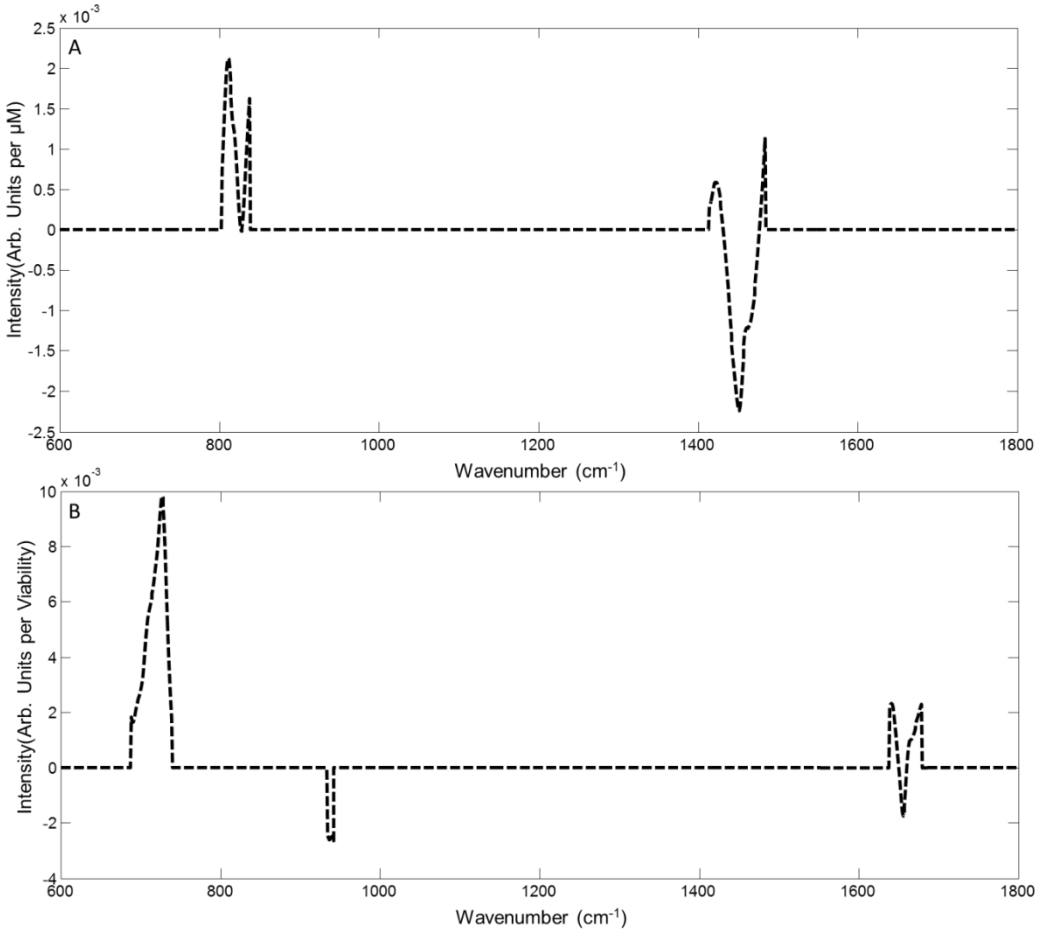


Figure 4.1: Spectral Constructs based on the normalised difference spectra between control and exposed nucleus (A) [10,] *and cytoplasm* [11] *(B). Selected Raman peaks were used to avoid over complexity in the simulated data; (A) the A form peak of DNA at 807 cm^{-1} and the B form peak at 833 cm^{-1} and the C-H deformation at 1449 cm^{-1} (B) the amide 1 band at ~1661 cm^{-1}, the C-C stretch intensity at ~939 cm^{-1} and the tryptophan peak at 731 cm^{-1}.*

4.3.4 Simulated data

Simulated datasets were generated in the following manner. A control dataset containing 25 spectra acquired from the nucleus of non-cis-platin exposed (control) cells was selected from Nawaz *et al* [1] (Figure 4.2). Notably, this real experimental dataset contains instrumental noise and sample variability. To this dataset, weighted contributions of the Concentration construct shown in Figure 4.1A, based on the experimentally observed difference spectra of the nuclear

region, were added, over the Lethal Concentration range 0.05 µM - 50µM used in the original study, based on a direct weighting of the spectral construct by the range of concentrations (Table 4.1). Initially, only the concentration dependent weighted constructs were added to the control, to produce Dataset 1.

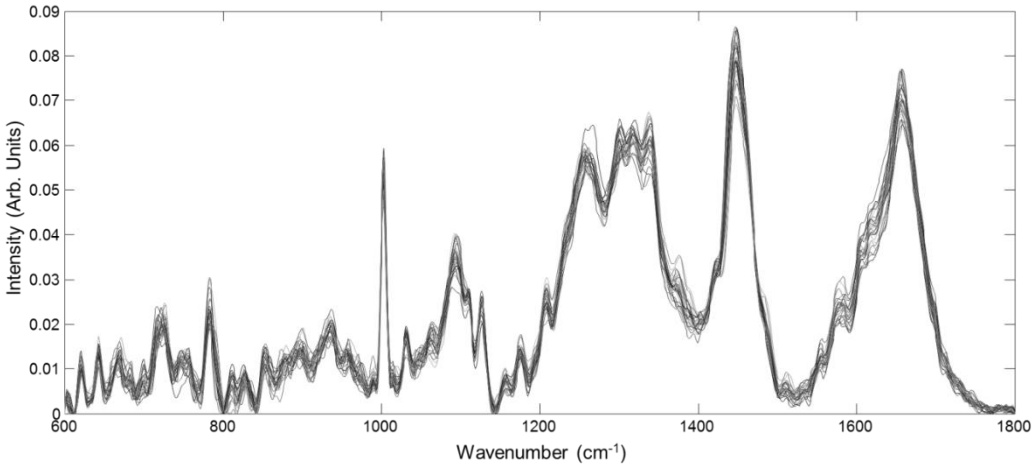


Figure 4.2: Control dataset taken from Nawaz et al.[1]; 25 control spectra taken from the nucleus of cells not exposed to cis-platin. Spectra have been baseline corrected and vector normalised. The inherent spectral variability in the data is representative of real experimental conditions. These spectra were then used in the construction of 3 simulated datasets, each containing 8 different dose/viability points with systematically introduced variation of the spectral constructs shown in figure 4.1.

As the MTT assay is expressed in viability compared to control (0.845 being maximum (V_{max}) and 0.135 being minimum values of fit to the experimentally observed viability over the concentration range[10]), the spectral construct of Figure 4.1B, derived from the experimentally observed differences in the cytoplasmic region, was similarly weighted by the (V_{max} – MTT) endpoints in Table 4.1 and also added to Dataset 1. Each spectral construct was therefore added following a linear trend based on concentration (Figure 4.1A) plus a linear

trend based on MTT response (Figure 4.1B). The MTT endpoint data are, however, nonlinearly related to the concentration, in a sigmoidal fashion typical of cytotoxic responses, as shown in Nawaz *et al*[1,2]. The resultant dataset therefore contains 25 spectra for each of 8 dose points (including control) which incorporate spectral variations, systematically dependent on both the exposure dose and the measured cytological response. For simplicity, this is referred to as Dataset 2.

It is noted that the spectral construct of Figure 4.1B is derived from exposure dose dependent, experimentally observed, spectral changes in the cytoplasmic region. No direct biological significance is inferred by the weighted addition of this spectral construct to the dataset derived from the nuclear regions. However, the addition serves to provide an independently variable perturbation to the dataset, which may serve to mimic a cytological effect of the direct action of the drug in the nucleus.

To probe the sensitivity of the methodology, the experimental range for cis-platin (Lethal Concentration, in table 4.1) has been extended (Sub lethal Concentration in table 4.1) to represent non-lethal doses of the drug. The MTT values have also been extrapolated according to the original fit of the Hill equation[10] to reflect these changes in concentration (Sub-lethal MTT in table 4.1). The corresponding simulated dataset will be referred to as Dataset 3. A dataset was also constructed which consisted solely of control spectra. This Control dataset did not contain any systematically introduced spectral variations and was used to establish a baseline regression endpoint for both Lethal Concentration and Lethal MTT.

Lethal Concentration	Sub-lethal Concentration	Lethal MTT	Sub-lethal MTT
0.05	0.0005	0	0.000001
0.5	0.005	0.15	0.000001
1	0.01	0.35	0.000001
3	0.03	0.52	0.00001
5	0.05	0.55	0.0001
10	0.1	0.65	0.001
50	0.5	0.66	0.01

Table 4.1: The weightings of the spectral constructs added to the control data. The Lethal Concentration and Lethal MTT ranges are derived from the actual experiment data of references [1,2]*. Lethal MTT represents the values obtained when the experimental MTT value is subtracted from* V_{max}*. The Sub-lethal Concentrations extend the concentration range and are representative of sub-lethal doses of cis-platin, for which sub-lethal MTT values are derived from the extrapolated fit of the Hill equation in Reference 1.*

4.4 Results

4.4.1 Concentration Simulated data

The PLSR method aims to establish a model that relates the variations of the spectral data to a series of relevant targets. In this case, the spectral data is a series of simulated datasets which are based on known introduced perturbations based on cis-platin-cellular interactions as described in the previous sections.

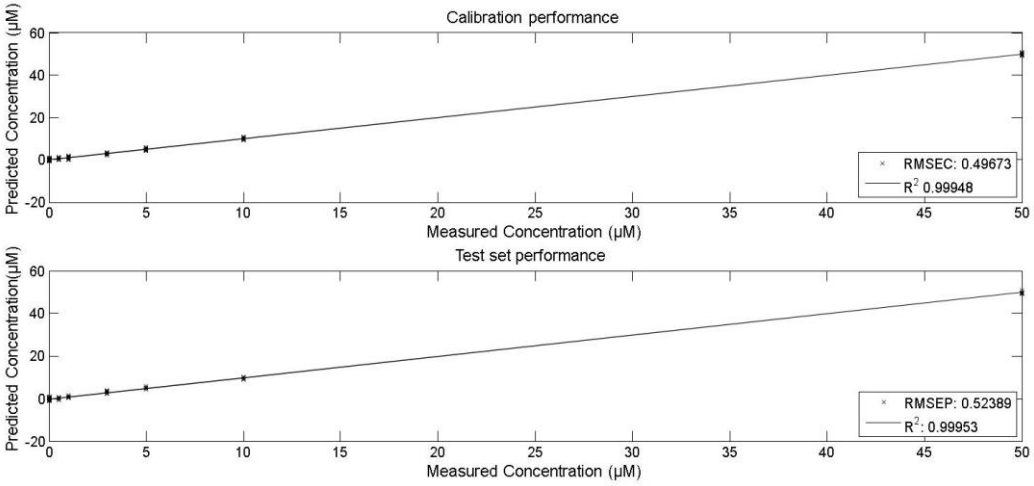


Figure 4.3. PLSR modelling against Lethal Concentration for Dataset 1. Top panel shows the calibration performance and test dataset (RMSEC 0.49673, R^2 0.99948). Bottom panel shows the performance of the model for the test dataset (RMSEP 0.52389, R^2 0.99953). Data was split in a ratio of 60:40 calibration and test respectively.

Regression of Dataset 1 against the Lethal Concentration range (table 4.1) yielded the model shown in figure 4.3. The data were split, 60:40, to create calibration and test sets to build the model. 60% of the data was used to calibrate the model and 40% of the data was then used to assess the performance of the model in predicting the expected target with unseen data. Leave-one out cross validation with the calibration set was used to determine the optimal model complexity for use in testing (Meade et al., 2010)[27]. This process was performed with randomization of the data matrix and splitting of the data to prevent data bias (Varmuza and Filzmoser, 2009)[28]. Control of over fitting was achieved using a procedure previously described by Martens and Naes[29]. The procedure involves selection of the optimal number of latent variables (LV) to retain within the PLSR model via cross-validation with the calibration data set. The optimal number of LV's was then selected on the basis of the number which provided the lowest root

mean squared error after cross validation. This is illustrated in Supplementary Material figure S4.1A and B, which show plots of the RMSECV and RMSEP for the first 10 LV's for the regression of Dataset 1 against Lethal Concentration 1, and thus the optimum number of LV's was selected as 10. The calibration and test set had RMSEC=0.49673, RMSEP=0.52389 and R^2 values of 0.99948 and 0.99953 respectively, indicating a good linear fit of the model.

As the regression co-efficient (RC) are descriptors of the spectral features which are used to build the model, we also aimed to assess the accuracy with which the algorithm can faithfully extract the known spectral perturbations introduced in the dataset. For regression of Dataset 1 against Lethal Concentration, we expect that the spectrum of the RC will be comprised of the Concentration construct which has been added based on the Lethal Concentration range (Figure 4.1A).

In figure 4.4, a direct comparison between the RC of regression of Dataset 1 against the Lethal Concentration range and the concentration spectral construct is shown. The spectrum of the RC is dominated by the peaks of the systematically added spectral construct, at 807cm^{-1}, 833cm^{-1} , which correspond to A and B form DNA[10] and the C-H deformation at 1449cm^{-1} (solid line figure 4.4 bottom panel). This verifies that the simulated changes are the major contributors to the PLSR model construction.

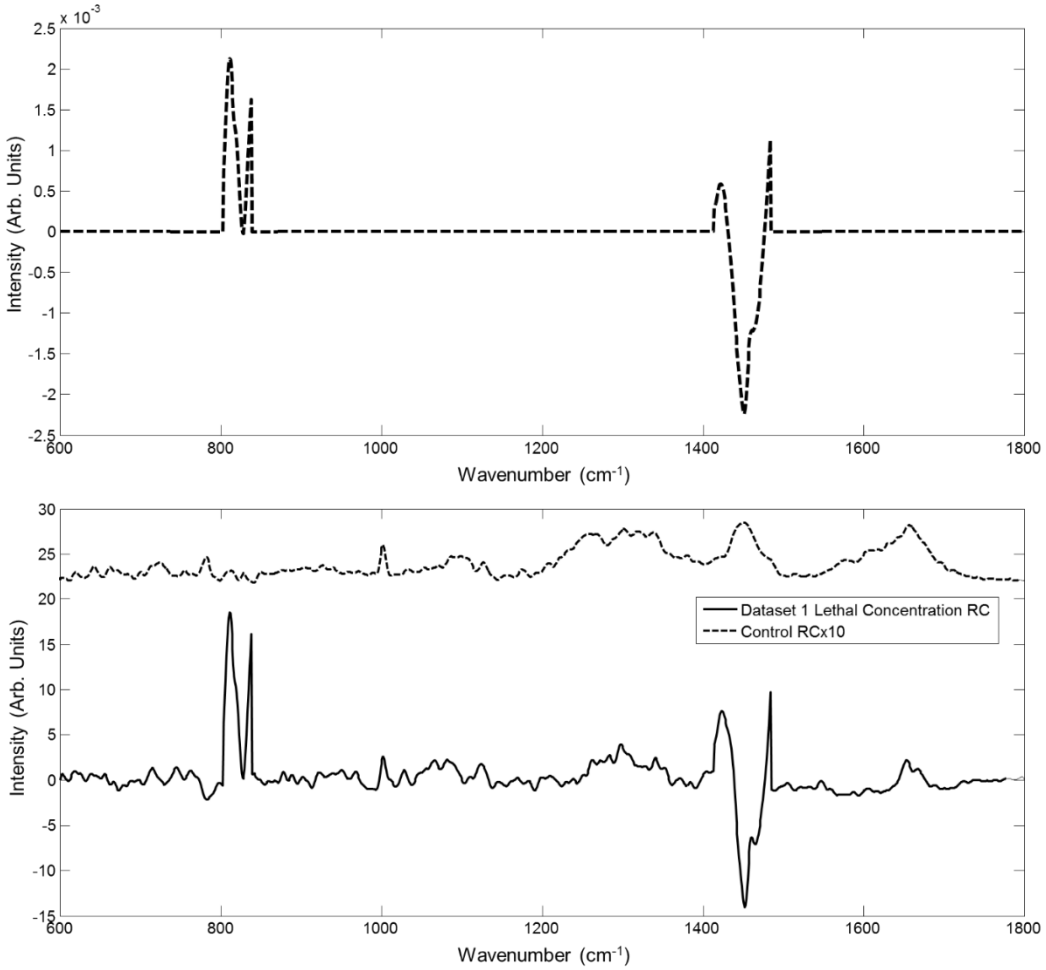


Figure 4.4: Plot of the regression co-efficient following PLSR of Dataset 1 against Lethal Concentration. The Concentration construct (dashed line) is shown in the top panel for comparison with the RC's in the bottom panel. The solid line (bottom panel) shows the regression co-efficient following regression of Dataset 1 against Lethal Concentration. The dotted line shows a plot of the regression co-efficient following regression of a dataset consisting of just control spectra against Lethal Concentration, in effect showing the baseline regression co-efficient when no introduced spectral perturbation (not including sample/instrumental variations) is present. The Control RC has been offset and multiplied by a factor of 10 for clarity.

However, it should be noted that the RC spectrum in figure 4.4 also contains other peaks which are not present in the spectral construct and so should not show a systematic variation with concentration. By regression of just the control data (with no spectral perturbations) against the Y target (Lethal Concentration) it was

possible to establish a Control RC, as shown by the dotted line (bottom panel) in figure 4.4 (offset and multiplied by a factor of 10 for clarity). The control RC spectrum shows a high degree of similarity with the original cellular spectra (Figure 4.2) and thus derives from the inherent variability in the experimental measurement. Close examination of the RC for the Dataset 1 regression reveals that some of the peaks in the Control RC are also present.

The PLSR modelling process was repeated for Dataset 2, which included the combined perturbations of the Concentration construct of Figure 4.1A, linearly weighted according to Lethal Concentration of Table 4.1, and the MTT Construct of Figure 4.1B, linearly weighted according to Lethal MTT of Table 4.1. A similar performance of model calibration and test were achieved, with RMSEC=0.4981, RMSEP=0.53505 and R^2 values of 0.99947 and 0.99952 respectively, again indicating a good linear fit of the model (Figure S4.2). The spectrum of RC again faithfully reproduced the Concentration Construct of Figure 4.1A, on a background which matches well the Control RC spectrum (Figure S4.3).

4.4.2 MTT Simulated Data

Dataset 2 also contains systematic perturbations which have been weighted according to the viability as measured using the MTT assay, and it is of critical interest whether these spectral variations can be independently extracted using PLSR, as suggested by Nawaz et al.[10]. Regression of Dataset 2 against Lethal MTT (table 4.1) yielded the model shown in figure 4.5. As for the concentration dependent model, the data are split according to 60% calibration and 40% test data. The calibration and test set had RMSEC=0.10158, RMSEP=0.12087 and R^2

values of 0.91928 and 0.89793 respectively. Based on these values, it can be seen that, while the model has fitted the data, it does not provide as good prediction as shown for concentration (figure 4.3). This is also reflected by the lower R^2 values, considering that the accuracy of the linear fit is measured by how close the value is to 1. A possible explanation for this is the lower magnitude and range of weightings of spectral construct added corresponding to the MTT response (Table 4.1, Lethal MTT).

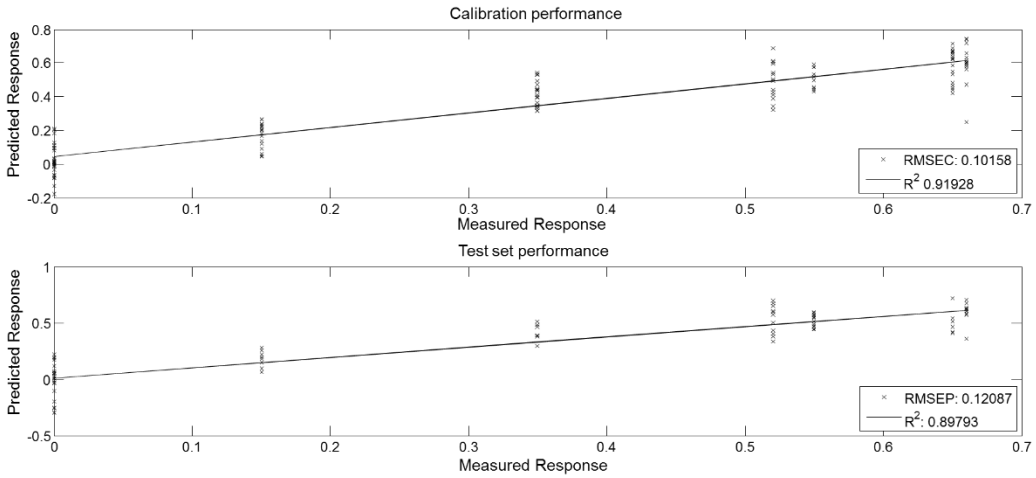


Figure 4.5: PLSR modelling of Dataset 2 against the Lethal MTT target. Top panel shows the calibration performance and test dataset (RMSEC 0.10158, R^2 0.91928). Bottom panel shows the performance of the model for the test dataset (RMSEP 0.12087, R^2 0.89793). Data has been split in a ratio of 60:40 calibration and test respectively.

Inspection of the MTT RC in Figure 4.6 shows that the peaks of the systematically added Viability construct (Figure 4.6, dashed line, top panel), the amide 1 band at ~1661 cm^{-1}, the C-C stretch intensity at ~939 cm^{-1} and the tryptophan peak at 731 cm^{-1}, are faithfully reproduced and dominate the MTT RC (Figure 4.6, solid line, bottom panel).

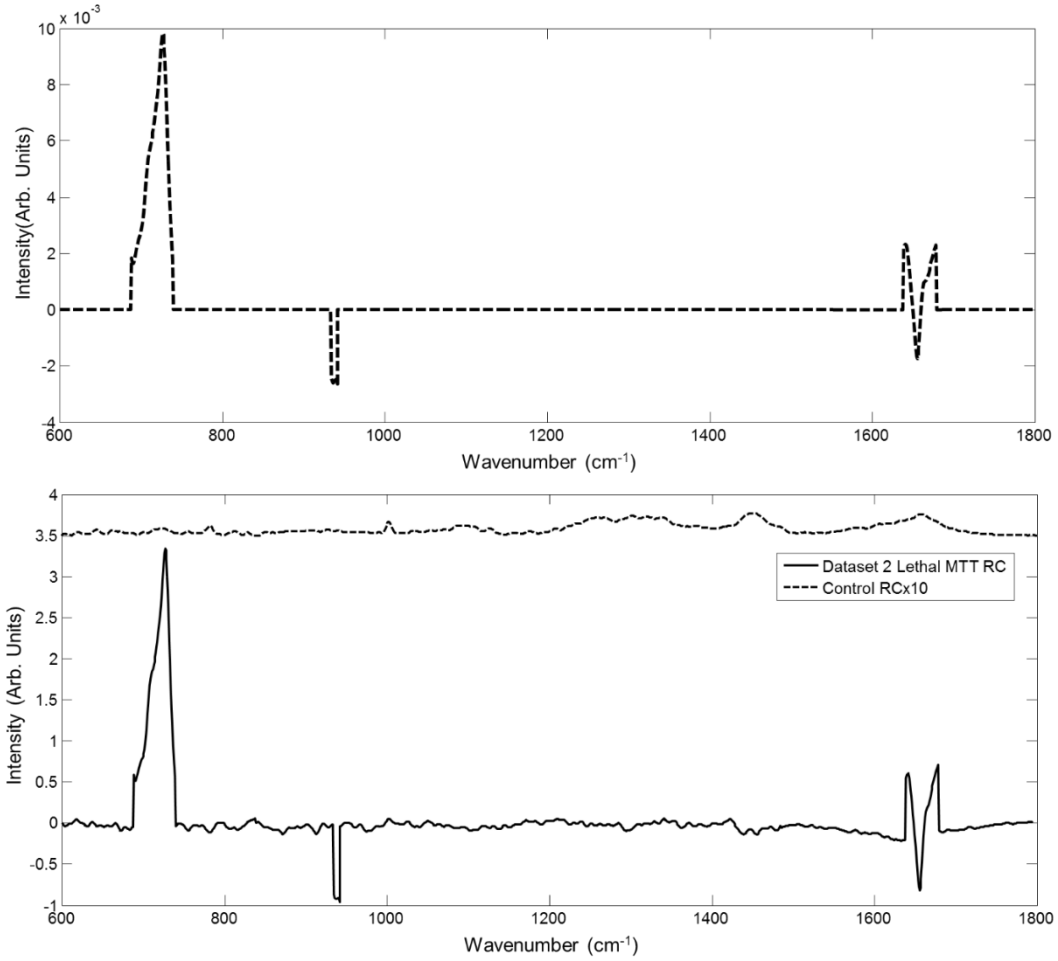


Figure 4.6: Plot of the regression co-efficient following PLSR modelling against MTT response. The Viability construct (dashed line) is shown in the top panel for comparison with the RC's in the bottom panel. The solid line shows the regression co-efficient following regression against Lethal MTT and Dataset 2 (bottom panel). The dotted line (bottom panel) shows a plot of the regression co-efficient following regression of a dataset consisting of just control spectra against Lethal MTT, in effect showing the baseline regression co-efficient when no introduced spectral perturbation (not including sample/instrumental variations) is present. The Control RC is offset and multiplied by a factor of 10 for clarity.

The baseline sensitivity is evaluated by regressing the control dataset against the Lethal MTT target, yielding the Control RC of Figure 4.6 (bottom panel, dotted line). The resultant RC spectrum has been offset and multiplied by a factor of 10, for clarity. As in the case for regression against Lethal

Concentration targets, the Control RC resembles the cellular spectra of figure 4.2, indicating that the baseline variation is limited by the variations in the original spectral measurement.

4.4.3 Quantitative evaluation of regression co-efficient

In an attempt to evaluate the quantitative nature of the regression co-efficient, a method was devised which looked at varying the number of data-points used to build the PLSR model. For the analysis of the spectral variations of Dataset 1, based on variations of the Concentration construct of figure 4.1A weighted according to Lethal Concentration (Table 4.1), multiple regressions were conducted (models not shown). Each model was constructed by increasing the number of data points, C+1 being the first data set used, consisting of the control dataset (Fig 4.2) and the 0.05 µM data-point of the Lethal Concentration range (Table 4.1). The data set was then successively extended by 1 data-point, such that C+2 consists of control, 0.05 µM and 0.5 µM, and so on, until all data points in the Lethal Concentration were included.

For all models, the spectrum of the RC displayed a combination of the Concentration construct of Figure 4.1A and the Control RC of Figure 4.4, and, as expected, regression over the full range reproduced the RC spectrum of Figure 4.4. Notably, as shown in Figure 4.7, the peaks of the Concentration construct increase linearly as the range of the regression is increased and reach a saturation value above ~ C+4. Extension of the model to 1000µM results in no further significant increase of these maximum peak intensities (data not shown). The A-form DNA peak at 807cm^{-1} reaches a maximum value of 18.46. Although this does not quantitatively equate to the corresponding peak value of the Control

construct of Figure 4.1A, the relative magnitudes of the respective peaks is consistent with those of the original Concentration construct, and notably the relative contribution of the Control RC is reduced with increasing range.

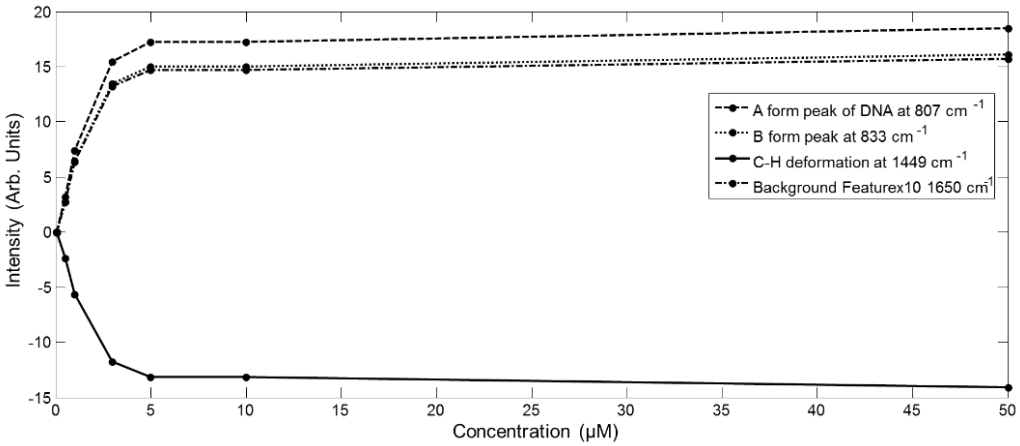


Figure 4.7: Evolution of the peaks of Construction construct in PLSR models of increasing range for Dataset 1.

A similar analysis was conducted for the PLSR of Dataset 2 against the Lethal Concentration range. Figure 4.8 shows a plot of the extracted RCs for all successive regressions. As expected, C+7 reproduces the Lethal Concentration RC of Figure 4.4, and extracts the expected introduced spectral construct (Figure 4.1 A). However, notably for all other regressions, C+1 to C+6, the presence of peaks which are not explicitly dependant on Lethal Concentration are observed. In addition to those of the Control RC, peaks of the MTT construct (Figure 4.1B) are evident in the RCs of the regressions over the incomplete concentration range. A similar phenomenon can be seen in the equivalent sequential modelling of the MTT data of Dataset 2 (Figure S4.4 and S4.5).

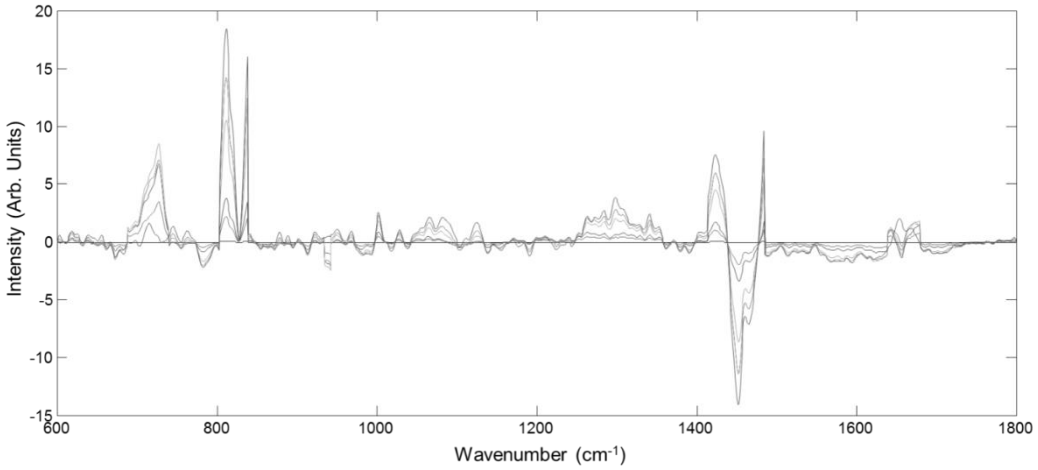


Figure 4.8. A plot of regression co-efficient following multiple regression against concentration with increasing data points. I.e. C+1 represents a dataset consisting of the control dataset and the data point at 0.05 µM. This then increases C+n until all data points in the dataset have been evaluated.

Figure 4.9 shows a plot of selected RC peak intensities associated with the spectral construct relating to concentration following successive rounds of regression as described above, namely the A form peak of DNA at 807 cm^{-1} and the B form peak at 833 cm^{-1}, which are associated with the physical changes associated with cis-platin-cellular interaction[2]. In fact the evolution of the peaks is observed to be identical to that observed for Dataset 1, shown in Figure 4.7, and although the plot of Figure 4.9 is in a linear/logarithmic format, it can be seen that the predicted relative intensities again increase linearly initially, before reaching a point of saturation at, or above, the dataset C+4, and further addition of data-points makes no difference (data not shown) to the quantitative prediction of the features.

Also shown in Figure 4.9 is the dependence of the peak of the Viability construct at 731cm^{-1}, (for example) which "bleeds through" in the regression of Dataset 2 against the incomplete concentration range. This bleed through occurs

for all peaks of the MTT Construct. The contribution of the peaks of the Viability Construct follows a trend of the derivative of the viability curve, indicating that it is the rate of change of the contributed spectral variations which governs the contribution to the RC. Notably, when the full Lethal Concentration range is included in the model, at the extremes of which the change in viability has reduced to the minimum value, the bleed through of the MTT construct is minimal, and the Concentration Construct of Figure 4.1A is faithfully extracted, albeit with an underlying background as a result of the inherent spectral variability.

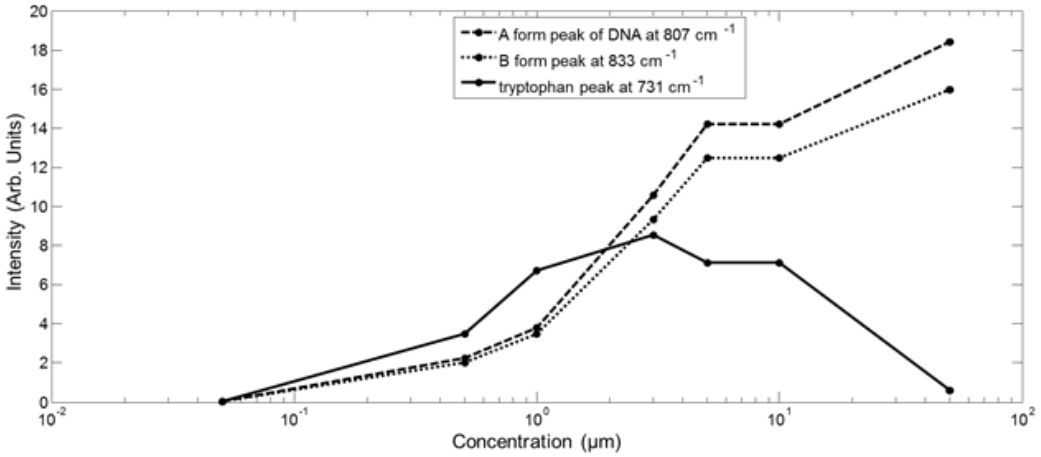


Figure 4.9. Plot of peak intensities vs. concentration of regression co-efficients for the A form peak of DNA at 807 cm^{-1} and the B form peak at 833 cm^{-1} of the Concentration Construct (Figure 4.1A). Also plotted is the contribution of the tryptophan peak at 731cm^{-1}, a key feature of the Viability Construct (Figure 4.1B)

A similar PLSRA of the contributions of the Viability construct to Dataset 2 reveals similar bleed through and more complex evolution of the features contributing to the spectrum of the RC (Supplementary Material Figures S4.4 and S4.5). The bleed through of the features of the spectral constructs shown in Figures 4.8 and 4.9 is a clear demonstration that it is not trivial to independently

extract the contributions of the two constructs over the lethal concentration range, as speculated by Nawaz *et al.*[10]. However, over concentration ranges in which the viability does not change significantly, the bleed through is minimal, and the concentration dependent spectral changes can be independently extracted. Thus, it should be possible to determine the direct chemical interactions of an external agent in the sub-lethal range.

Figure S4.6 shows the calibration and test performance of the PLSR of Dataset 3 versus the sub-lethal concentration range of Table 4.1. The model yields RMSEC and RMSEP values of 0.143 and 0.19392, respectively, with R^2 values of 0.38916 and -0.24063, accuracies considerably less that those of the equivalent model in the Lethal Concentration range. Notably, the RC spectrum is a faithful extraction of the pure Concentration construct of Figure 4.1 A, as shown in Figure S4.7. Little or no bleed through of features associated with the Viability construct is apparent (although still present in minimal quantities) although this is not surprising as, with little or no change in viability, the contributions of the Viability construct to Dataset 3 are minimal.

4.5 Discussion

Given the drive for a reduction in the use of animal models for evaluating toxicity, screening of drugs and even cosmetics, due to regulatory developments in both the EU and US (EU Directive-2010/63/EU and US Public Law 106-545, 2010, 106th Congress)[30–32] generally based on the 3 R's of Russell and Burch[30] to replace, reduce and refine the use of animals used for scientific purposes, there is increased emphasis on the development of reliable and rapid *in-vitro* screening methodologies. This includes more representative culture models which better

mimic the *in-vivo* environment as well as more rapid, cost efficient, high content, and ideally label free screening technologies. It is crucial, however, that these models and technologies are well validated against established gold standards[33,34].

Raman spectra, in principle, contain high content information about the biochemical make up of the sample, and changes to it, related to pathology or an external agent. Raman spectra contain numerous peaks which vary dependently and independently of each other. Crucially, for real applications and particularly in the instance of drug interactions, it is difficult to tell whether these differences are inherently based on cell to cell variability or whether they are dependent on the primary action of the drug (i.e. the direct chemical effects) or the secondary effects the drug has on the cell (i.e. the response of the cell to said drug).

In this study, simulated datasets were used to evaluate the capability of PLSR to extract known and systematic spectral variation from a control dataset, which contained intrinsic experimental variability. The spectral variations introduced varied linearly with the applied drug dose and also with the measured cell population response, as measured by a standard cytotoxicity assay. Notably, however, the two spectral variations are not completely independent, as the viability response is sigmoidal dependent on the applied dose.

In the case where only a concentration dependent systematic variation in the spectra is introduced, the PLSR model provides an accurate predictive response tool, the regression co-efficient of which are based on the systematic variation which has been introduced to the dataset, linearly dependent on the targets. The model shows high sensitivity, and the limits of detection are determined only by the intrinsic variability of the experimental method, as determined by the PLSR of the Control spectral dataset. This limit can be

improved by optimising sample preparation and measurement protocols. In principle, such a PLSR model can predict the response of a drug dose in a cell population, or determine an unknown drug dose from a measured spectral response.

However, the spectral changes which result from the interaction and action of a drug within a cell are manifold, and it is of interest to differentiate the spectral signatures of the direct interaction from the subsequent cellular response. Notably, this study demonstrates that, although PLSR predictive models based on regression of the combined dataset, including all spectral responses, against the target of concentration range produce a similarly accurate, linear predictive model, the contributing RCs are only derived exclusively from the introduced concentration dependent variations in ranges where all other spectral variations are limited. For example, as shown in Figures 4.8 and 4.9, regression over the limited range of C+4 produces a model which is based on RCs which includes contributions derived from the direct effect of the interaction of the drug within the cell (Concentration construct), as well as the resultant cytological response (Viability construct). Thus, care should be taken in interpreting the spectral features which contribute to such regressions to elucidate the underlying mechanisms.

Nevertheless, in sub-lethal regions, the direct effects of the drug interaction can confidently be investigated employing such a PLSR analysis of Raman spectral data, independent of the cytological responses, and these are easily discernible above the intrinsic variability of the control. Although this seems a trivial conclusion, such rapid, label free analysis could prove invaluable in screening of, for example, the mechanisms and efficacy of drug interactions,

evaluating drug uptake and receptor binding[25] or nanoparticle uptake and trafficking in regions where cytoxicity assays are insensitive.

The use of a parallel cytotoxic assay such as MTT serves as a range finding test to establish the IC_{50}, but also provides vital information about the sub-lethal doses and maximum responses. It also provides a target for regression of the data in the regions of toxicity. Thus, the subsequent cytological effects can be differentiated from the direct chemical effects of the agent and extracted from the overall spectral response in the dose range where the viability is impacted, and the cellular response can be independently mapped spectroscopically, as a function of dose and time. Notably, the model described here, which includes a single spectral construct to represent the cellular response is very simplistic, as the response is a cascade of many responses, depending on the mechanism of interaction[35]. Nevertheless, the analysis presented here demonstrates that the spectral fingerprints of the direct mechanisms of interaction and the subsequent cellular responses can be independently extracted from the dose dependent spectral data, and thus, ultimately with improved screening sensitivities and speeds, Raman spectroscopy could be employed to monitor in quasi real time, in a label free manner, the efficacy and mode of action of, for example chemotherapeutic agents and other exogenous agents, laying the basis for improved quantitative structure activity relationships to guide drug development or chemical regulation strategies.

4.6 Conclusions

This study demonstrates the reliability and also limitations of PLSR as a method for predictive modelling and analysis of spectroscopic signatures of cellular

responses to exogenous agents such as radiation, chemotherapeutic agents or toxins. The spectroscopic profiles at any dose/time point can derive from a complex mixture of direct interactions within the cell and a cascade of subsequent cellular response. The analysis demonstrates that care should be taken in choosing the response range and also highlights the importance of parallel cytological assays in guiding the modelling and analysis. Correct choice of range can help differentiate between the signatures of direct interactions, which are dominant at sub-lethal doses and those of the subsequent cellular response which evolve with increasing dose.

The study also demonstrates the importance of simulated datasets in exploring the potential as well as the limits of the analytical techniques. Notably, the use of real experimental data which contains sample variability and instrumental response factors as a basis of the simulated dataset helps to visualise the lower limits of sensitivity.

The results indicate that Raman spectroscopic screening combined with such regression models and feature selection techniques, in parallel with conventional cytotoxicity assays, can be used to screen for the efficacy of drug interactions and can contribute to understanding the mechanisms of interaction.

4.7Acknowledgement

This research was supported by the Integrated NanoScience Platform, Ireland (INSPIRE), funded under the Higher Education Authority PRTLI (Programme for Research in Third Level Institutions) Cycle 5, co-funded by the Irish

Government and the European Union Structural fund, and Science Foundation Ireland (08/PI/11).

4.8 References

1 H. Nawaz, F. Bonnier, P. Knief, O. Howe, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2010, **135**, 3070–6.

2 H. Nawaz, F. Bonnier, A. D. Meade, F. M. Lyng and H. J. Byrne, *Analyst*, 2011, **136**, 2450–63.

3 F. M. Lyng, E. O. Faoláin, J. Conroy, a D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan and H. J. Byrne, *Exp. Mol. Pathol.*, 2007, **82**, 121–9.

4 I. Taleb, G. Thiéfin, C. Gobinet, V. Untereiner, B. Bernard-Chabert, A. Heurgué, C. Truntzer, P. Hillon, M. Manfait, P. Ducoroy and G. D. Sockalingum, *Analyst*, 2013, **138**, 4006–14.

5 P. Crow, B. Barrass, C. Kendall, M. Hart-Prieto, M. Wright, R. Persad and N. Stone, *Br. J. Cancer*, 2005, **92**, 2166–70.

6 T. J. Harvey, E. Gazi, A. Henderson, R. D. Snook, N. W. Clarke, M. Brown and P. Gardner, *Analyst*, 2009, **134**, 1083–91.

7 F. Bonnier, P. Knief, B. Lim, a D. Meade, J. Dorney, K. Bhattacharya, F. M. Lyng and H. J. Byrne, *Analyst*, 2010, **135**, 3169–77.

8 K. Klein, A. M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F. Jamitzky, G. Morfill, R. W. Stark and J. Schlegel, *Biophys. J.*, 2012, **102**, 360–8.

9 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

10 C. Matthäus, T. Chernenko, J. A. Newmark, C. M. Warner and M. Diem, *Biophys. J.*, 2007, **93**, 668–73.

11 P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J. H. Shanks, M. Brown, N. W. Clarke and P. Gardner, *Analyst*, 2012, **137**, 1370–7.

12 T. Chernenko, R. R. Sawant, M. Miljkovic, L. Quintero, M. Diem and V. Torchilin, *Mol. Pharm.*, 2012, **9**, 930–6.

13 J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers and H. J. Byrne, *Analyst*, 2012, **137**, 1111–9.

14 M. E. Keating, F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 5792–802.

15 P. Bassan, A. Kohler, H. Martens, J. Lee, H. J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke and P. Gardner, *Analyst*, 2010, **135**, 268–77.

16 H. Byrne, K. Ostrowska and H. Nawaz, *Opt. Spectrosc. Comput. Methods Biol. Med.*, 2014, **14**, 355–399.

17 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

18 A. D. Meade, H. J. Byrne and F. M. Lyng, *Mutat. Res.*, 2010, **704**, 108–14.

19 K. M. Ostrowska, A. Malkin, A. Meade, J. O'Leary, C. Martin, C. Spillane, H. J. Byrne and F. M. Lyng, *Analyst*, 2010, **135**, 3087–93.

20 R. M. Balabin and S. V. Smirnov, *Anal. Chim. Acta*, 2011, **692**, 63–72.

21 M. Jimenez-Hernandez, C. Hughes, P. Bassan, F. Ball, M. D. Brown, N. W. Clarke and P. Gardner, *Analyst*, 2013, **138**, 3957–66.

22 K. W. C. Poon, F. M. Lyng, P. Knief, O. Howe, A. D. Meade, J. F. Curtin, H. J. Byrne and J. Vaughan, *Analyst*, 2012, **137**, 1807–14.

23 H. Nawaz, A. Garcia, A. D. Meade, F. M. Lyng and H. J. Byrne, *Analyst*, 2013, **138**, 6177–84.

24 D. Rohleder, W. Kiefer and W. Petrich, *Analyst*, 2004, **129**, 906–11.

25 J. Black and P. Leff, *Proc R Soc L. B Biol Sci.*, 1983, **220**, 141–162.

26 S. Wold, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. …*, 2001, 109–130.

27 A. D. Meade, C. Clarke, H. J. Byrne and F. M. Lyng, *Radiat. Res.*, 2010, **2**, 225–37.

28 K. Vermuza and P. Flizmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics.pdf*, CRC Press, 2009.

29 H. Martens and T. Næs, *Multivariate Calibration.pdf*, John Wiley & Sons, 1994.

30 THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION, *Off. J. Eur. Union*, 2010, 33–79.

31 U. S. Congress, 2001, 2721–2725.

32 W. Russell, R. Burch and C. Hume, *The principles of humane experimental technique*, Methuen, London, 1959.

33 A. Tfayli, F. Bonnier, Z. Farhane, D. Libong, H. J. Byrne and A. Baillet-Guffroy, *Exp. Dermatol.*, 2014, **23**, 441–3.

34 F. Bonnier, M. Keating, T. Wróbel, K. Majzner, M. Baranska, A. Garcia, A. Blanco and H. J. Byrne, *Toxicol. Vitr.*, 2014, **29**, 124–131.

35 M. A. Maher, P. C. Naha, S. P. Mukherjee and H. J. Byrne, *Toxicol. Vitr.*, 2014, **28**, 1449–60.

4.9 Supplemental Material:

Multivariate statistical methodologies applied in biomedical Raman spectroscopy: Assessing the validity of partial least squares regression using simulated model datasets.

Mark E. Keating[1,2]*, Haq Nawaz[3] , Franck Bonnier[1,4] and Hugh J. Byrne[1]
[1]FOCAS Research Institute, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland.
[2]School of Physics, Dublin Institute of Technology, Kevin Street, Dublin 8 Ireland.
[3]National Institute for Biotechnology and Genetic Engineering (NIBGE), P.O.Box 577, Jhang Road Faisalabad, Pakistan.
[4]Faculty of Pharmacy, EA 6295 – NM/NP, Université François-Rabelais de Tours, 60 rue du Plat D'Etain, 37020 Tours Cedex 1, France

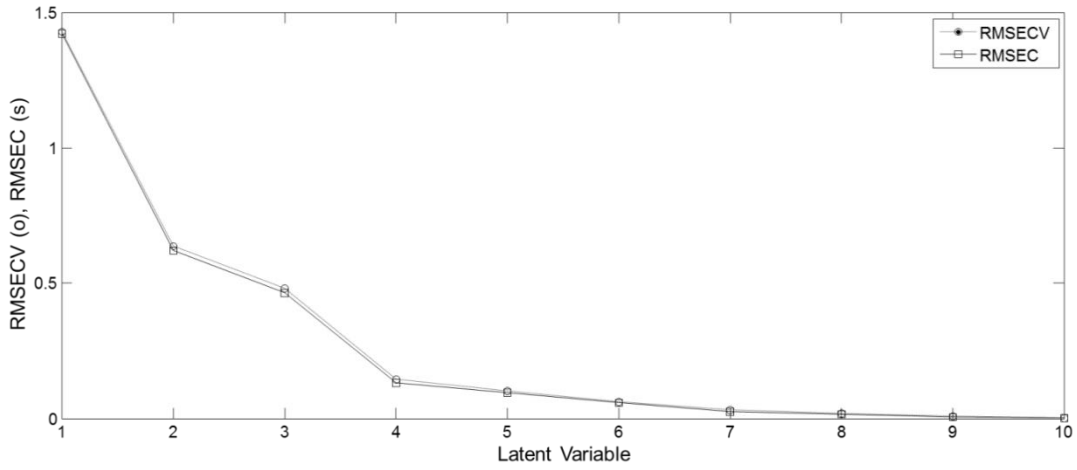


Figure S4.1: RMSECV and RMSEP for the first 10 LV's for the regression of Dataset 1 against Lethal Concentration 1

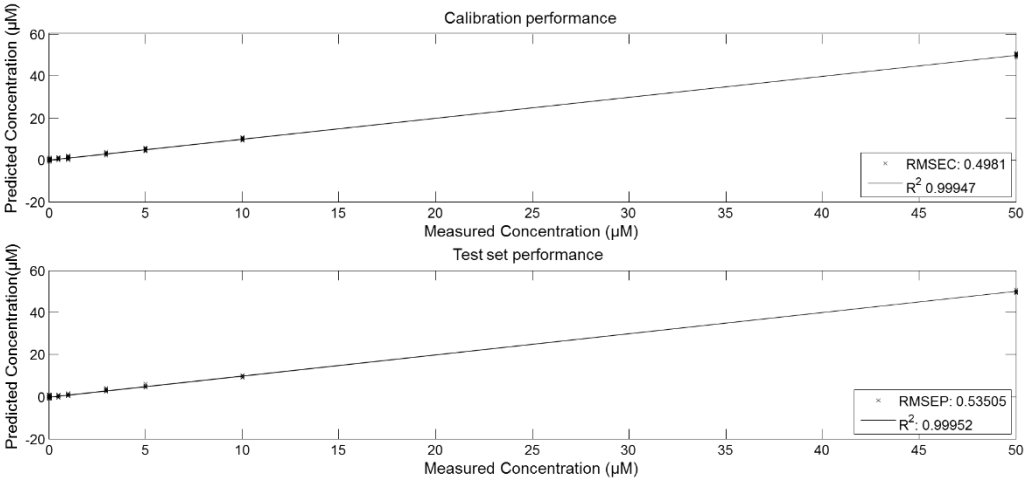


Figure S4.2: PLSR modelling of Dataset 2 with the Lethal Concentration range as taregt. Top panel shows the calibration performance and test dataset (RMSEC 0.4981, R^2 0.99947). Bottom panel shows the performance of the model for the test dataset (RMSEP 0.53505, R^2 0.99952). Data was split in a ratio of 60:40 calibration and test respectively.

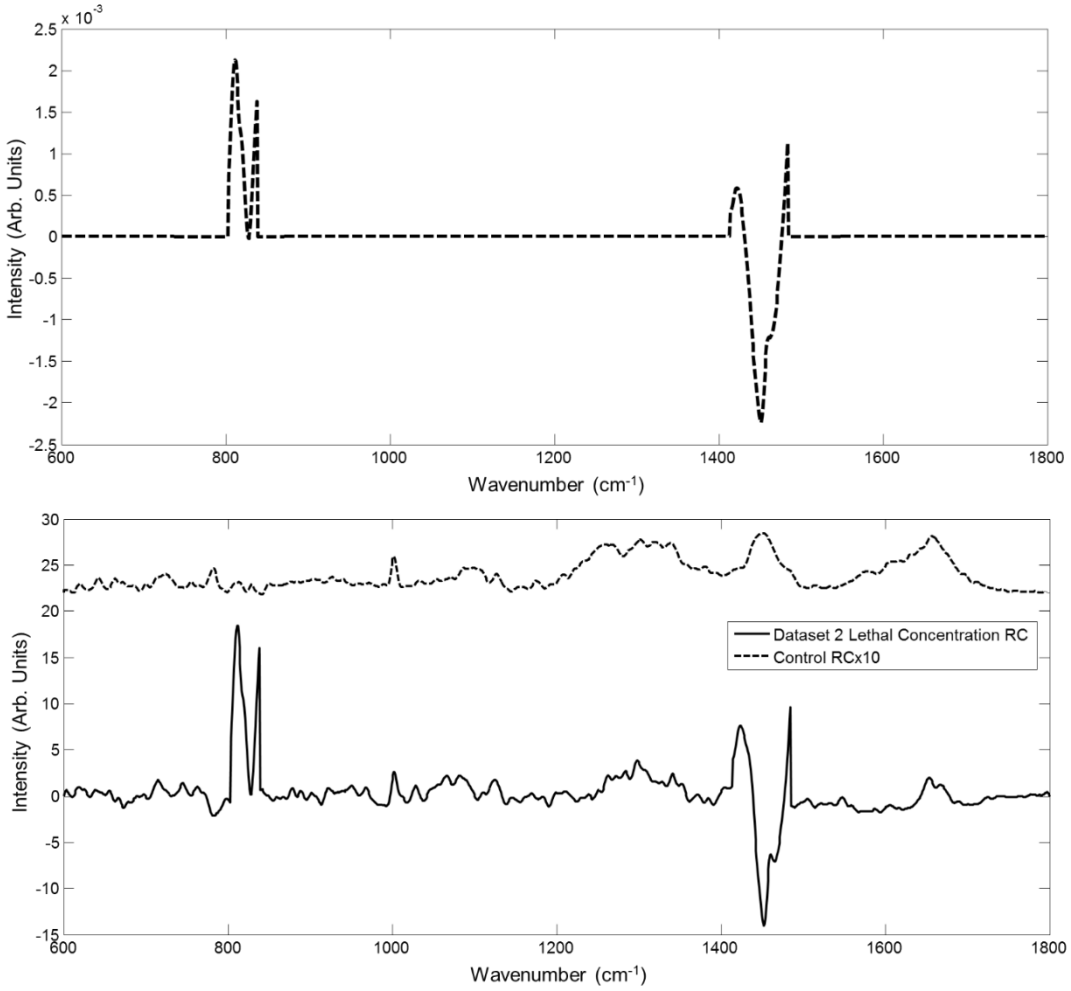


Figure S4.3: Plot of the regression co-efficient following PLSR modelling of Dataset 2 against Lethal Concentration. The concentration spectral construct (dashed line) is shown in the top panel for comparison with the RC's in the bottom panel. The dashed line (bottom panel) shows the spectrum of regression co-efficients following regression of Dataset 2 against Lethal Concentration 1. The solid line shows a plot of the regression co-efficient following regression of a dataset consisting of just control spectra against Lethal Concentration, in effect showing the baseline regression co-efficient when no introduced spectral perturbation (not including sample/instrumental variations) is present. The Control RC has been multiplied by a factor of 10 and offset for clarity.

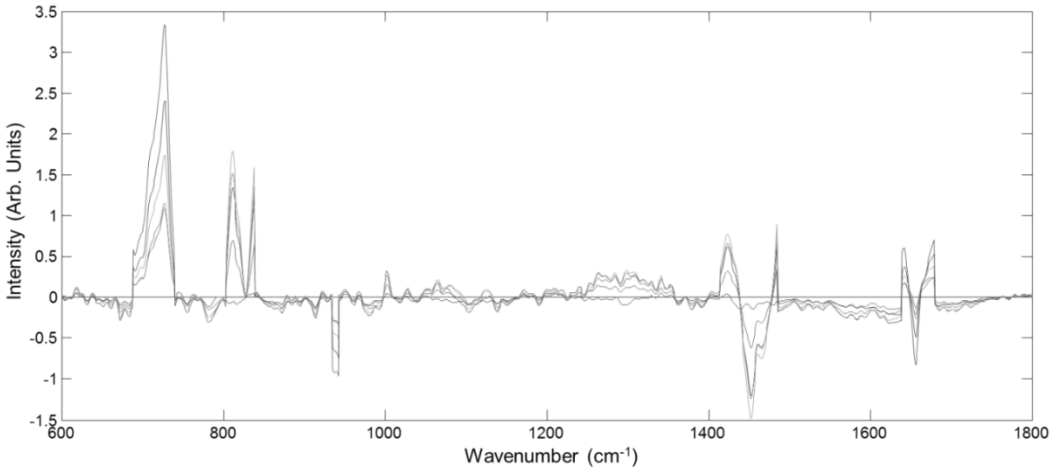


Figure S4.4. A plot of regression co-efficients following multiple regression of Dataset 2 against Lethal MTT with increasing data points. I.e. C+1 represents a dataset consisting of the control dataset and the data point at 0.05 µM. This then increases C+n until all data points in the dataset have been included.

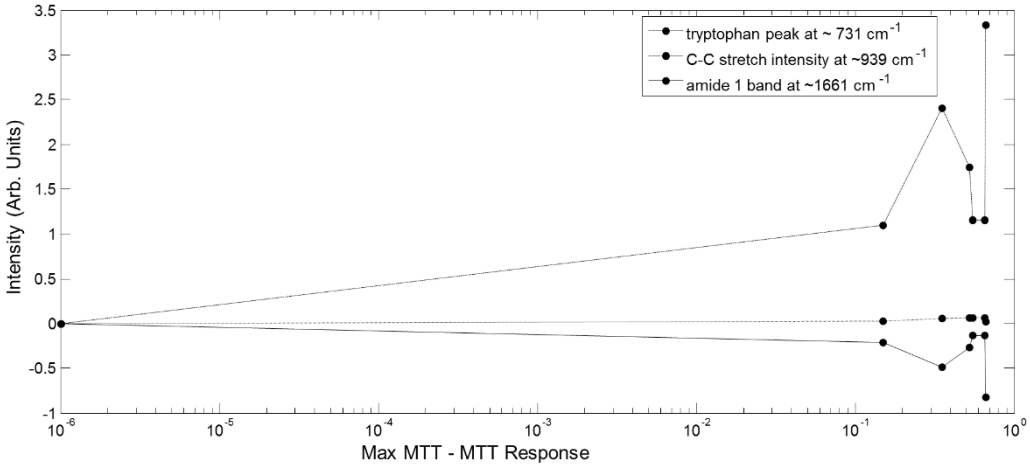


Figure S4.5 Plot of RC peak intensities for regression of Dataset 2 against Lethal MTT; C-C stretch intensity at ~939 cm^{-1}, the amide 1 band at ~1661 cm^{-1} and the tryptophan peak at 731 cm^{-1} of the Viability Construct (Figure 1B).

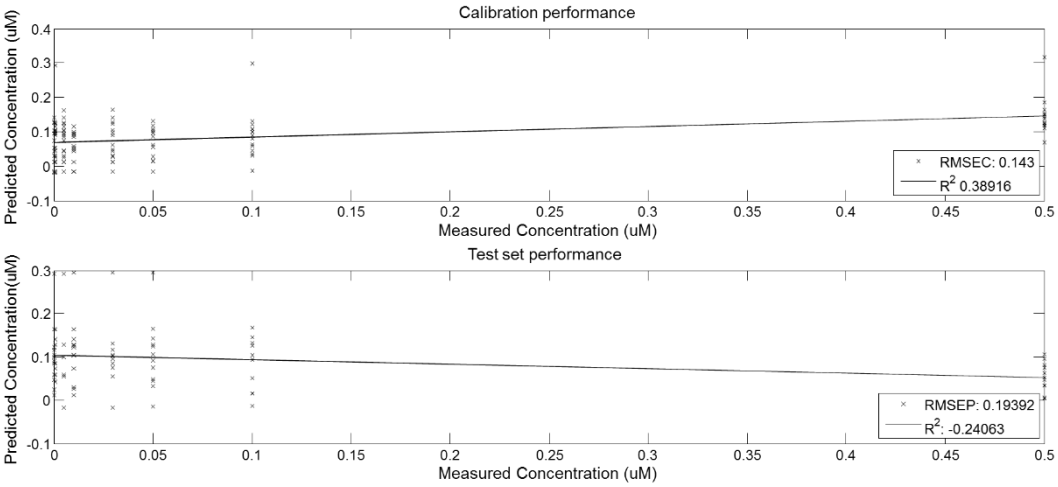


Figure S4.6. PLSR modelling of Dataset 3 with the Sub-lethal Concentration range as target. Top panel shows the calibration performance and test dataset (RMSEC 0.143, R2 0.38916). Bottom panel shows the performance of the model for the test dataset (RMSEP 0.19392, R2 -0.24063). Data was split in a ratio of 60:40 calibration and test respectively.

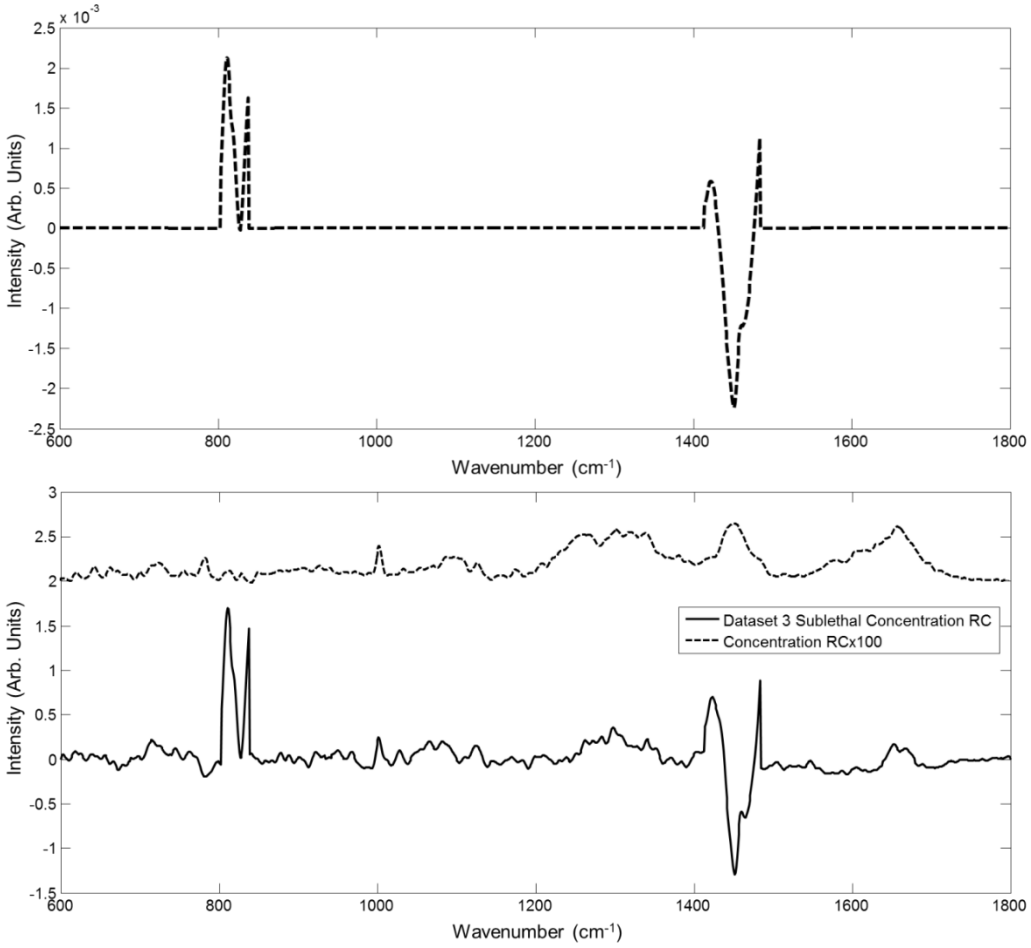


Figure S4.7. Plot of the regression co-efficients following PLSR of Dataset 3 against Sub-lethal Concentration. The concentration spectral construct (dashed line) is shown in the top panel for comparison with the RC's in the bottom panel. The solid line shows the regression co-efficient following regression against sub-lethal concentration and Dataset 3 (bottom panel). The dotted line (bottom panel) shows a plot of the regression co-efficient following regression of a dataset consisting of just control spectra against sub-lethal concentration, in effect showing the baseline regression co-efficient when no introduced spectral perturbation (not including sample/instrumental variations) is present. The Control RC is offset and multiplied by a factor of 100 for clarity.

Chapter 5 Seeded Principal Component Analysis for biochemical screening using vibrational spectroscopy

5.1 Abstract:

Seeded Principal Component Analysis (SePCA) is introduced as a novel multivariate analysis variant to address some of the limitations of the application of PCA in bio-spectroscopy in systematically varying datasets. Using simulated data based on experimental spectra of *in-vitro* exposure to varying doses of the chemotherapeutic agent, cis-platin, standard and SePCA are compared, firstly based on their ability to differentiate the responses to different exposure doses, and secondly to assess the accuracy of the loadings that are used to describe the systematic variations of biochemistry underlying the differentiation. Further insights are also garnered on the use of 1st and 2nd derivative spectra and the impact this mathematical transformation has on the ability of the algorithm to separate and describe the spectral origin of differentiation of spectral datasets. The implications of this novel variant of PCA are discussed in the context of screening for drug efficacy *in-vitro* as well as biomedical classification for disease diagnostics.

5.2 Introduction

Raman spectroscopy is a branch of vibrational spectroscopy which allows for a sample to be characterised based on its inherent chemical nature in a label free manner. The resulting spectra can then be used to classify a host of materials from organic to inorganic compounds. Recently, vibrational spectroscopies such as IR

and Raman have gained momentum in a biomedical context, with a guiding focus of translating these technologies into the clinical environment[1].

Both histopathological and cytological studies have been routinely carried out using both IR and Raman spectroscopic analysis and have shown significant success in the laboratory at diagnosing disease states with high sensitivity and specificity, based on the inherent biochemical state of the sample as opposed to the morphology[2]. Further studies have investigated other disease states such as atherosclerosis[3], liver cancer and disease[4–6],lung cancer[7], colon cancer[8], blood borne illnesses such as malaria and others[9,10], as well as investigations in dermatology[11]. Additionally, bio-fluid analysis has also been developed in conjunction with IR and Raman spectroscopy as a potential alternative to current gold standard practices[12]. *In-vivo* and *ex-vivo* Raman spectroscopy has also been demonstrated in a number of studies, including the investigation of brain[13], cervical[14] and oesophageal[15] pathologies[16]

Vibrational spectroscopy has also seen usage in other medical contexts such as Nanomedicine[17], in which Raman and its variants allow for label free characterisation of nanomaterials in cells as well as tissues and live animal studies. Pharmacological characterisation *in-vitro* has also been demonstrated with a number of drugs classified using Raman spectroscopy as a tool to monitor drug behaviour in a cellular environment[18–22].

Crucially, there are a number of caveats associated with this technique and all play a role in the end goal of accurate label free sample characterisation. These range from correct sample preparation in the laboratory, precise sampling by the spectrometer (incorporating sample location as well as instrument precision and reproducibility), through to the correct spectral pre-processing,

including baseline correction, smoothing, normalisation…etc. Finally, the correct usage and development of novel multivariate statistical methodologies for analysis and interpretation of the biochemistry underlying the spectral signatures is important for validity, accuracy and interpretability[1,23,24].

In previous publications by Keating et al[25,26], simulated datasets were used to probe the intricacies of Partial Least Squares Regression (PLSR) and a novel multivariate approach, termed spectral cross correlation analysis (SCCA) and its use in biomedical Raman spectroscopy. In brief, the validity of the method to extract and differentiate the spectral signatures of the direct action of and subsequent metabolic response to chemotherapeutic agents and nanoparticles (respectively) in cells *in-vitro*, was demonstrated. The use of simulated datasets, based on real experimental data, enabled verification of the validity of the techniques and estimation of the limits of sensitivity, while also identifying potential limitations of the techniques, highlighting the importance of understanding the intricacies of multivariate statistical methodologies applied in vibrational spectroscopy.

In a study by Bonnier and Byrne in 2011, principal component analysis (PCA) was investigated and its use and interpretability in spectral applications was elucidated[27]. While this method is commonly applied in biomedical spectroscopy as a means of differentiating and classifying spectral datasets, the biochemical reasoning behind separation and its dependence on the loadings can often be misinterpreted. This publication aimed to shed light on some of the possible pit falls in interpreting spectral separation using both real and simulated data.

Following on from this work, as well as the validation of the PLSR technique by Keating et al[25] , the current study aims to further probe the PCA algorithm using simulated data, and in particular to explore its use to analyse systematic variations in spectral responses as a result of quasi-continuous variations in exposure of cell populations to external stimuli. Some of the deficiencies of the method are highlighted and investigated, looking at continuously varying data and loading interpretability in a spectral context that of systematic variations due to variable doses of chemotherapeutic agents in cells, *in-vitro*. A novel variant protocol for carrying out PCA, termed seeded PCA (SePCA), has been developed to overcome some of the deficiencies in the current usage of the algorithm in spectroscopy.

5.3 Methods

5.3.1 Simulated data

The generation of the simulated data used has been described in detail elsewhere[25]. A modified version of the protocol is described here to tailor the dataset for the study of PCA. The full details of the culture and experimental conditions can also be found elsewhere[18] In brief, human lung adenocarcinoma (A549) cells were routinely cultured at 37 °C, 5 % CO_2 in DMEM F12 supplemented with 10% FBS, 1% pen/strep and 2mM l-glutamine. Cells were cultured until 70-80% confluency and plated on quartz substrates for Raman spectroscopy. Twenty five control spectra were acquired in the previous study by Nawaz et al[18]. Figure 5.1 shows the control spectra used. These consist of spectra acquired from the nuclei of A549 human lung adenocarcinoma cells *in-vitro*, with

no external agent added. In parallel, a standard MTT cytotoxicity assay was carried out to establish the dose dependence of the in-vitro viability.

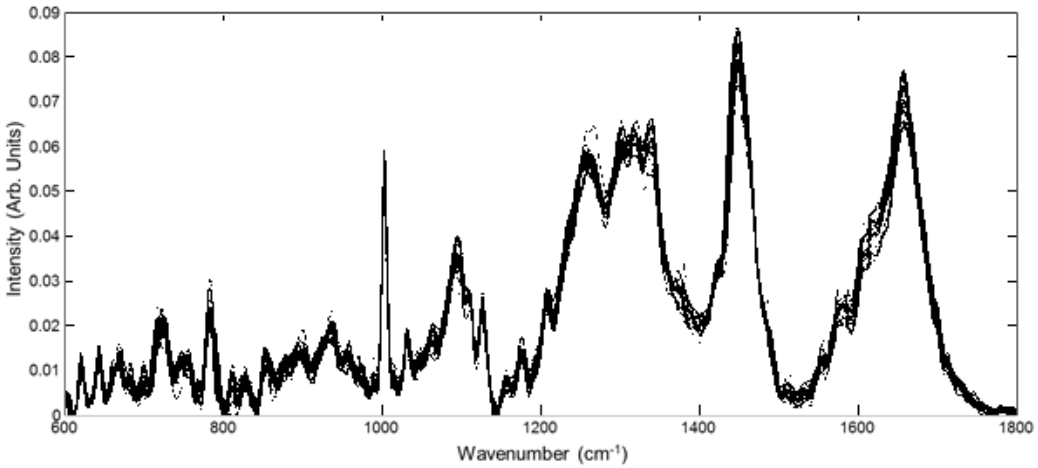


Figure 5.1. Control dataset taken from Nawaz et al[18]*. 25 control spectra taken from the nucleus of cells not exposed to cis-platin. Spectra have been baseline corrected and vector normalised. The inherent spectral variability in the data is representative of real experimental conditions.*

Spectral constructs, shown in figure 5.2, were generated from the mean difference between spectra from concentration dependant changes in the nucleus and the control, as well as spectra from the cytoplasm and the control, which correspond to the MTT cytotoxic response as described in the publications by Nawaz et al. [18,19]. In these publications, the exposure of cells to cis-platin was carried out over the concentration range from 0.05µM-50µM, including the mean inhibitory concentration, IC_{50} of ~ 3µM, as identified using the MTT cytotoxicity assay. This dose range in conjunction with the MTT response was then used to construct simulated data. In this way, spectral constructs were generated from the changes in the spectra of the nuclear region, including increases in the

characteristic A form of DNA peak at 807 cm^{-1} and the B form peak at 833 cm^{-1} and a change in the C-H deformation at 1449 cm^{-1} (Figure 5.2A) and in the cytoplasmic region, containing the following peak changes or shifts; a change in the amide 1 band at ~1661 cm^{-1}, a decrease in the C-C stretch intensity at ~939 cm^{-1} and an increase in the tryptophan peak at 731 cm^{-1} (Figure 5.1B).

These were then used to introduce systematically variable, known perturbations into the control dataset, in such a way as to generate a number of different datasets based on the original experimental results obtained by Nawaz et al[18]. to investigate both the standard PCA algorithm and the novel seeded variant.

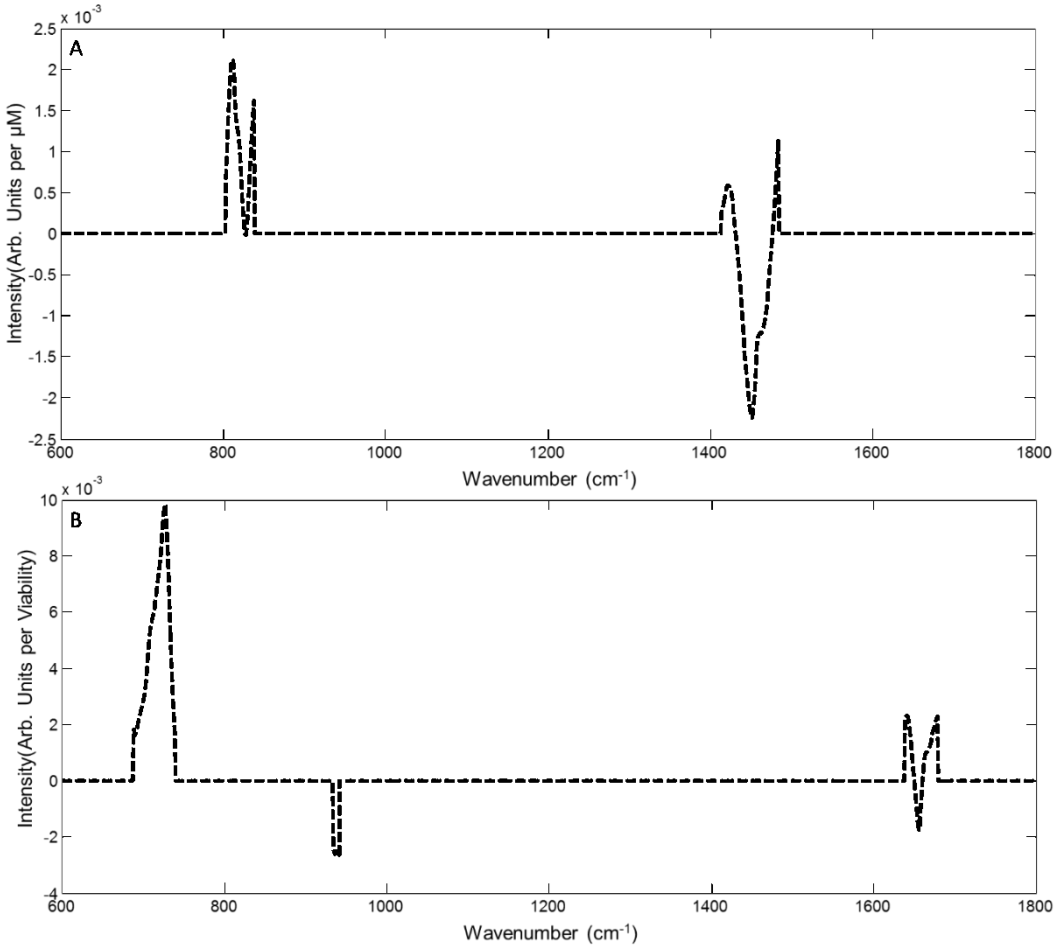


Figure 5.2: Spectral Constructs based on the normalised difference spectra between control and exposed nucleus (A) and cytoplasm (B) of Nawaz et al. (2010). Selected Raman peaks were used to avoid over complexity in the simulated data; (A) the A form peak of DNA at 807 cm^{-1} and the B form peak at 833 cm^{-1} and the C-H deformation at 1449 cm^{-1} (B) the amide 1 band at ~1661 cm^{-1}, the C-C stretch intensity at ~939 cm^{-1} and the tryptophan peak at 731 cm^{-1}.

The perturbation of construct 1 (Figure 5.2 (A)) is used to represent the systematic spectral evolution due to the direct interaction of the chemotherapeutic agent in the nucleus, while construct 2 (Figure 5.2 (B)) is used to represent the systematic variations in viability of the cell population, as measured by the MTT assay. The weightings listed in table 5.1 show the magnitude of the changes added to the control dataset of figure 5.1. Three datasets were generated from these sets of values, one of which consists of the control (25 spectra) plus the maximal MTT (25 spectra) value of table 5.1 (Dataset 1),

Dataset 1: (Control), (Control + $WMax_{MTT}$)

where $WMax_{MTT}$ indicates the addition of the maximally weighted MTT spectral construct in table 5.1 i.e. 0.66*MTT spectral construct (Construct 2).

The second dataset (Dataset 2) contains the control and the maximally weighted values for both the concentration (Construct 1) and MTT (Construct 2) constructs. Each of the weighted constructs, $WMax_{Conc}$ and $WMax_{MTT}$, are added to the control dataset separately in the first instance with the purpose of creating a simulated system to probe the ability of standard and seeded PCA to accurately separate the three groups, each containing 25 spectra, which make up the dataset.

Dataset 2: (Control), (Control + $WMax_{MTT}$), (Control + $WMax_{Conc}$)

A further, more complex version of the dataset was constructed using all the weightings in table 5.1, generating a dataset with simultaneous, continuously varying concentration and continuously varying MTT constructs, the final data matrix having the format of: 7x25 weighted spectra plus 25 controls.

Dataset 3: (Control), (Control +$W_{MTTn+Conc.n}$ +$W_{MTT.n+1+Conc.n+1……}$)

Dataset 3 thus simulates the spectral changes observed experimentally as a result of exposure of A549 cells *in-vitro* to cis-platin of continuously varying concentration.

Approximate partial first and second derivative spectra were calculated using the 'diff' function in Matlab (Mathworks, USA)), with seeded and standard PCA carried out on selected datasets (1-3) to highlight the improvements made possible by this mathematical transformation.

n	Concentration	MTT
1	0.05	0.1
2	0.5	0.15
3	1	0.35
4	3	0.52
5	5	0.55
6	10	0.65
7 (Max)	50	0.66

Table 5.1 the weightings of the spectral constructs added to the control data. The Concentration and MTT ranges are derived from the actual experiment data of references[18,19] *MTT represents the values obtained when the experimental MTT value is subtracted from the maximum viability*

5.3.2 PCA

PCA is a method which aims to reduce the dimensionality of the data to describe the variation present in a dataset, whereby the first principal component is a description of the maximum variance present in the dataset, the second describes the second most variance…etc. The principal component scores can then be described by the loading vector, which is a representation of this variance. In a Raman spectroscopy context, the scores represent values which correspond to a loading spectrum which contains peaks, both positive and negative, which explain the spectral variation in the dataset[17].This tool can be quite useful for classifying spectra into groups e.g. diseased and non-diseased[14]. It has also been used to reconstruct images[28,29], i.e. a variance plot based on the loadings plot. However, as these loadings plots may often contain a number of spectral features corresponding to different cellular biochemistry, interpretation can be difficult

and it is quite possible to misinterpret. Bonnier et al. have shown that pairwise PCA of clusters identified by KMCA can provide a clearer picture of the specific biochemical differences between region[30]. All spectra were centred using a single value decomposition (SVD) algorithm for analysis using the following function. All analysis was done using the 'pca' function in Matlab (The Mathworks Inc.).

5.3.3. Seeded PCA

Considering that PCA is sensitive to the relative scaling of the original variables, it is postulated that seeding the examined dataset with the known perturbation might have the effect of increasing the accuracy and or sensitivity of the standard PCA algorithm. In the case of interest, the spectral signature of the nuclear binding of the chemotherapeutic agent may be known, and thus the experimentally observed spectral changes may be data-mined for this signature. This is done by the addition of the pure spectral constructs of Figure 5.2 to the Datasets of control and perturbed control spectra described in Section 5.3.1.

To optimise for seeding using an optimised seeded weight (O_{SW}), it is necessary to understand how the variance changes as the magnitude of the spectral construct added to the dataset is increased. To do this, an exponentially increasing weighting was explored, as is shown in table 5.2. This involved systematically increasing the weighting of the spectral construct added to the dataset and examining how the explained variance changes as a result.

Multiplicative Weighting	Magnitude
Weighting 1	10^1
Weighting 2	10^2
Weighting 3	10^3
Weighting 4	10^4
Weighting 5	10^5
Weighting 6	10^6
Weighting 7	10^7
Weighting 8	10^8
Weighting 9	10^9
Weighting 10	10^{10}

Table 5.2: weightings used to multiply the spectral constructs of figure 5.2 for the determination of the optimum magnitude for seeded PCA.

5.4 Results

5.4.1. PCA Dataset 1

Figure 5.3 shows the scatter plot and loadings for PC 1, 2 and 3 following a standard PCA of the simplified dataset consisting of the control and simulated spectra based on the maximal MTT value, Dataset 1, shown in table 5.1. PC 1, PC 2 and PC3 describe the majority variance in the dataset, PC 1, 2 and 3 accounting for 37.98%, 24.36% and 6.58% of the explained variance, respectively. No discernible separation of the control and MTT dataset is represented by the scatter plot in figure 5.3A, and this is reflected by the loadings

shown in figure 5.4A and B. However, marginal partition of the data is demonstrated in figure 5.3B and C, although. no separation between control and maximal MTT data is demonstrated. This is due to the inability of PCA to extract out the ‘pure’ spectral features introduced in the data, which are not evident in the loadings of any of the first 3 PCs (Figure 5.4).

This poses a problem, as it is not possible to show with standard PCA a separation based on the introduction of a known spectral perturbation, or extract information concerning that perturbation, as the intrinsic spectral variability is larger than that of the systematic variability introduced. To address this, a seeded PCA (SePCA) methodology was developed, which allows for a ‘pure’ spectral loading, in this case the MTT spectral perturbation, to be extracted with the maximal variance described by the desired changes.

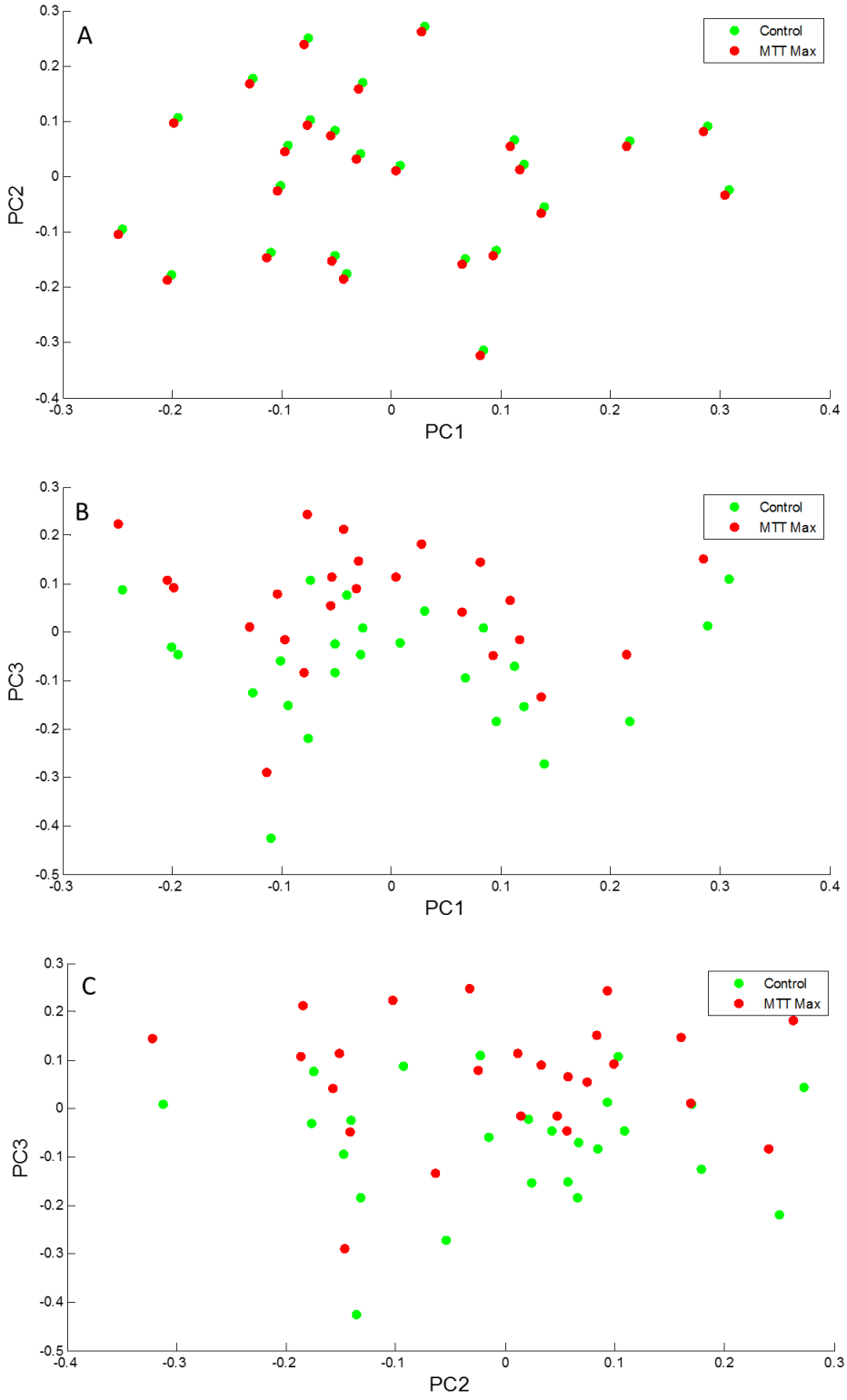


Figure 5.3. PCA on a dataset consisting of the control and max MTT, dataset 1 (A) scatter plot of PC1 vs. PC2 (B) scatter plot of PC1 vs. PC3 (C) scatter plot of PC2 vs. PC3. PC1, 2 and 3 account for 37.98%, 24.36% and 6.58% of the variance in the dataset, respectively.

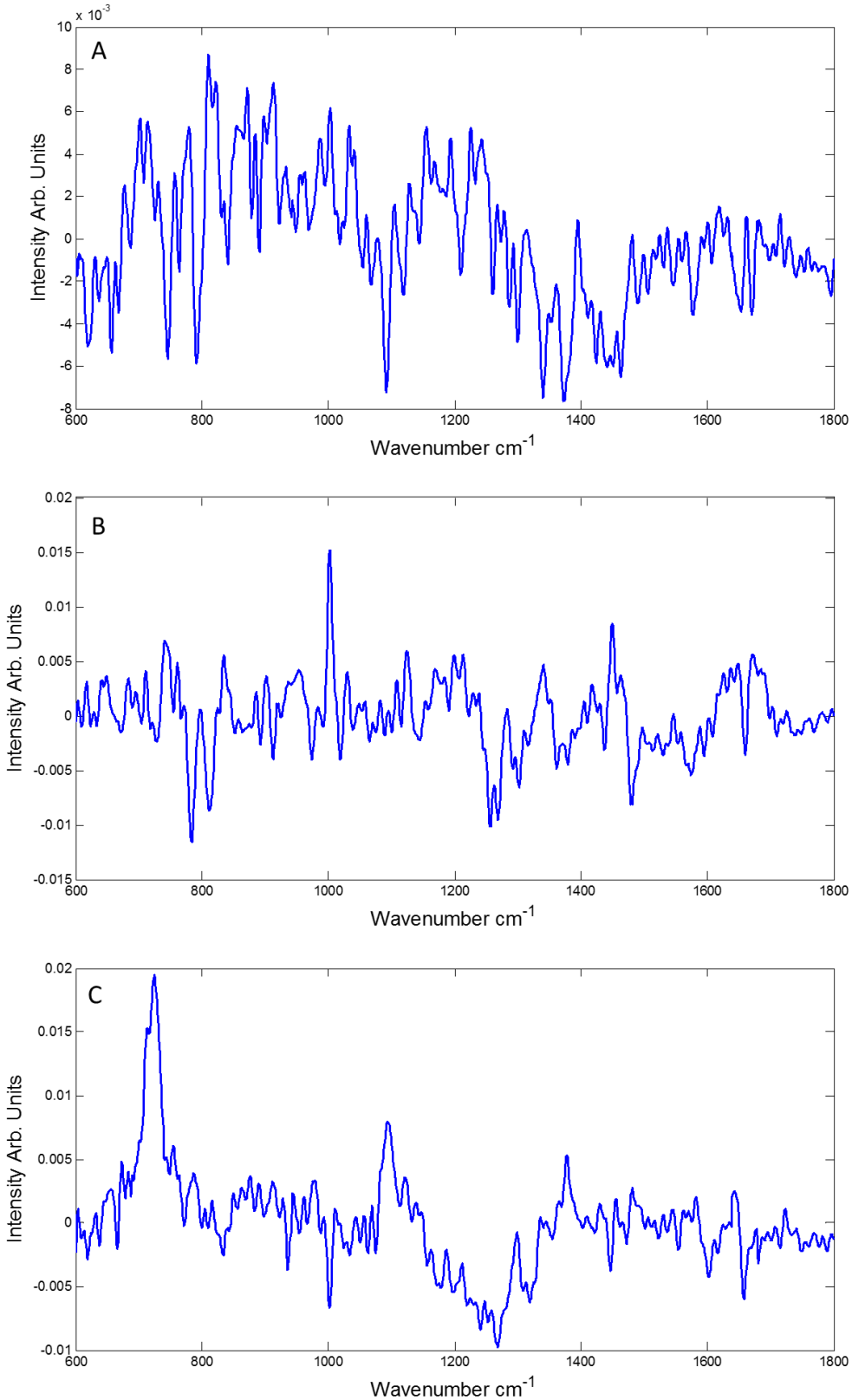


Figure 5.4. Loadings corresponding to (A) PC1, (B) PC2 and (C) PC3 for standard PCA on dataset 1. With PC 1, 2 and 3 accounting for 37.98%, 24.36% and 6.58% of the variance respectively.

5.4.2. Seeded optimisation

Seeding of the dataset was performed as described in Section 5.3.3. Seeded PCA. To optimise for seeding, the variance explained by PC1 and PC2 as the seeded weighting is increased according to table 5.2 is examined. The optimisation of the weight for seeding is shown in figure 5.5A and B, showing the increase in variance explained by the MTT seeded loading, and decreased variance explained by the intrinsic spectral variance. “Whereas at 0 weighting, PC1 is dominated by the intrinsic variance, at a seeding of 10^2, PC1 begins to show contributions of MTT, while at a loading of 10^3, it is almost completely dominated by features of Construct 2 and completely dominated by Construct 2 at 10^4 .Therefore, based on this plot, the optimal weight to seed for this dataset is ~10^4 with 99.99% of the variance explained by the first PC and an almost negligible 0.01% described by the inherent dataset variability.

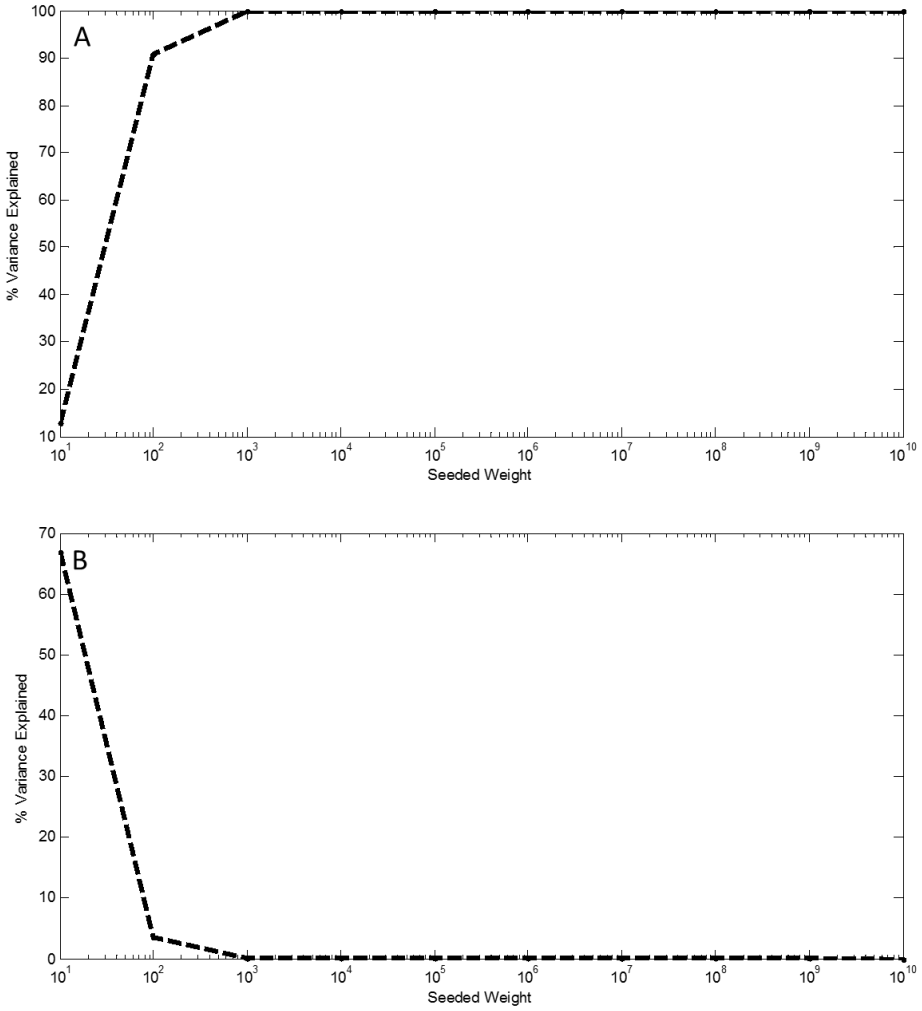


Figure 5.5: Calculation of the variance explained after successive rounds of PCA with increasing spectral construct weighting according to table 2. (A) % variance explained by the PC loading addition (B) % variance explained by the inherent dataset variability between control spectra acquired, instrumental error, random noise…etc

5.4.3. Seeded PCA dataset 1

Following seeding optimisation i.e. 10^4 weighted addition of the MTT spectral construct (Construct 2), it is possible to show partition of the data based on the known perturbations to the dataset. This is shown in Figure 5.6 A for Dataset 1, which now consists of the control spectra (25), the control + $WMax_{MTT}$ spectra (25) and $WMax_{MTT}x10^4$ (1). Separation between control and MTT is clearly demonstrated. The loadings of PC1, 2 and 3 are shown in figure 5.7 A, B and C and account for 99.99%, 0.000059% and 0.000038% of the variance respectively. The loading of PC1 faithfully reproduces the MTT spectral construct of Figure 5.2, the known spectral perturbation added to the dataset. The inherent variability of the dataset is now represented by PC2 and 3, and no partition of the data is evident along these axis. This demonstrates, in a simplified simulated dataset, the enhancement seeded PCA can achieve over the standard implementation of PCA. In real experimental data, however, the spectral variations are likely to be more complex, as for example in the study of Nawaz et al., in which the spectral variations as a result of a chemotherapeutic agent represented both the direct chemical effect of the drug as well as the subsequent cellular responses.

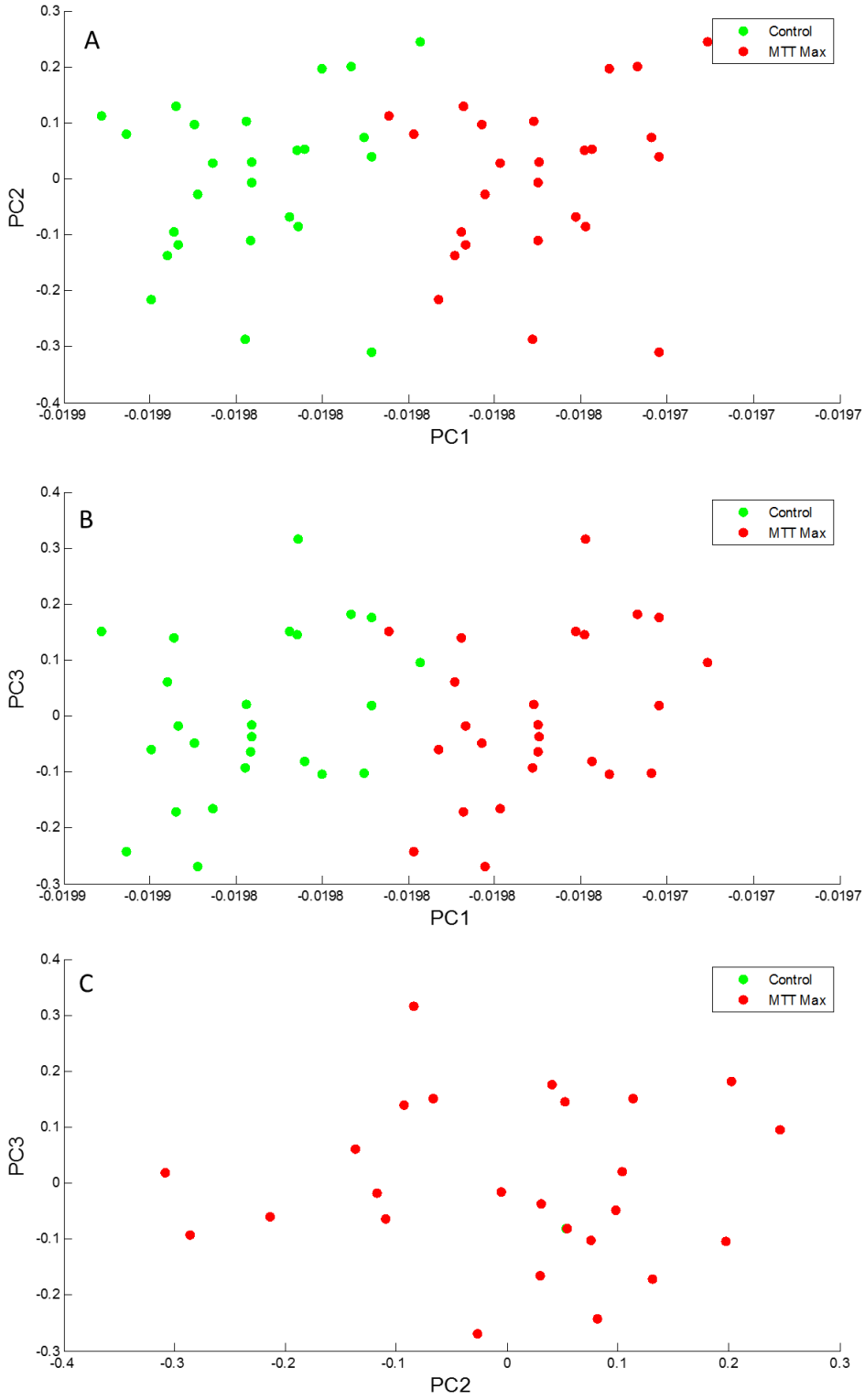


Figure 5.6. PCA on a dataset consisting of the control and max MTT, dataset 1 (A) scatter plot of PC1 vs. PC2 (B) scatter plot of PC1 vs. PC3 (C) scatter plot of PC2 vs. PC3. PC1, 2 and 3 account for 99.99%, 0.000059% and 0.000038% of the variance in the dataset, respectively.

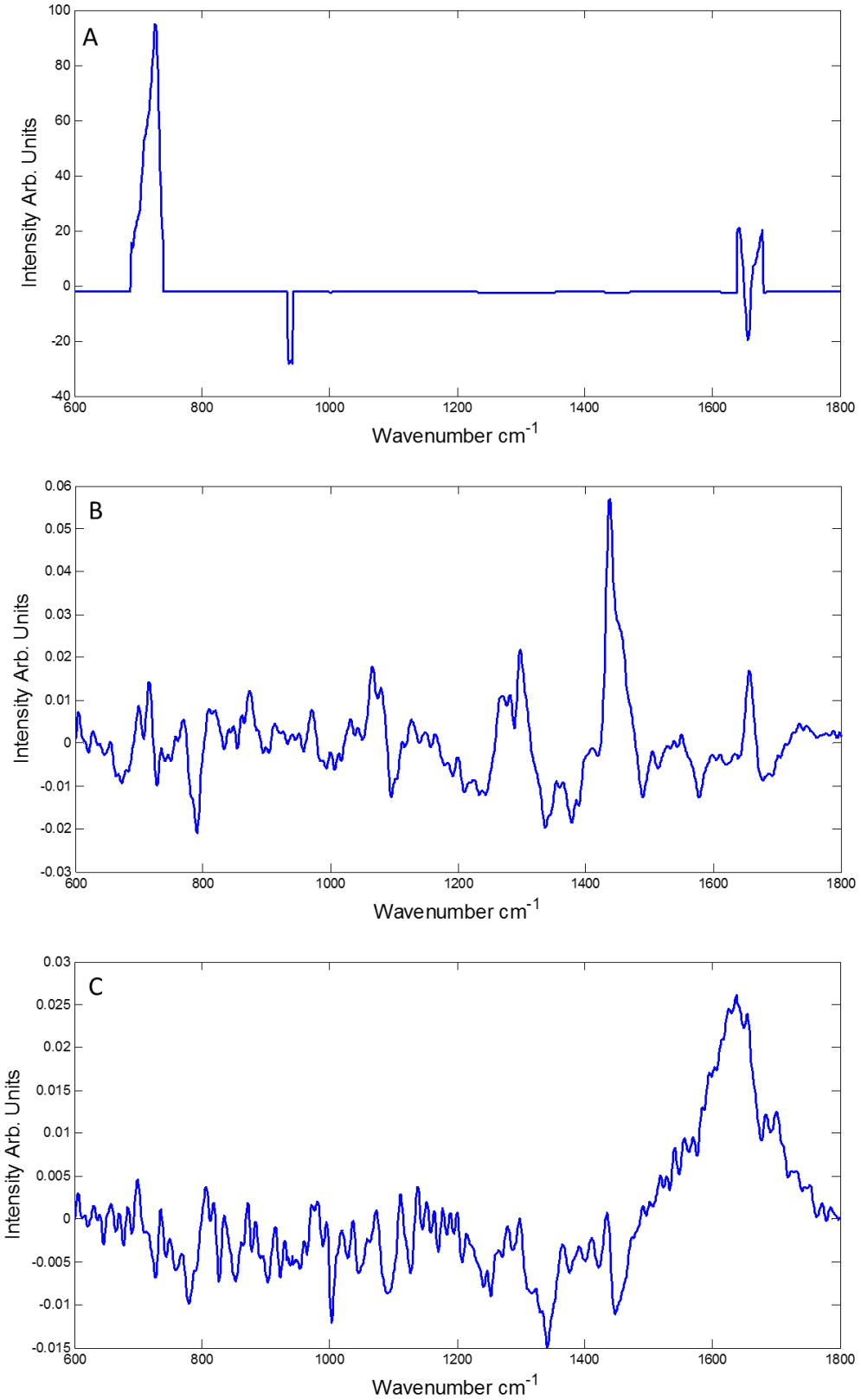


Figure 5.7. Loadings corresponding to (A) PC1, (B) PC2 and (C) PC3 for Seeded PCA on dataset 1. With PC 1, 2 and 3 accounting for 99.99%, 0.000059% and 0.000038% of the variance respectively.

5.4.4. PCA Dataset 2

As shown in Figure 5.8, PCA of dataset 2 results in partial differentiation according to PC1 and PC2, which account for 93.94%, 2.22% and 1.54% of variance in the dataset. PC1 differentiates the spectra perturbed by $WMax_{Conc}$ from those of control and perturbed by $WMax_{MTT}$. Spectral features pertaining to Construct 1 are present in the 1st principal component as shown in figure 5.9A.

Similar to the case of dataset 1, however, the spectral features of the MTT spectral construct introduced into the data are not shown in the first three PC's. Therefore, no separation between control and max MTT is shown in figure 5.8 A-C. Thus, while conventional PCA can extract the spectral features associated with the chemical interaction of the chemotherapeutic agent with the cell, it is not sensitive to the weaker changes associated with the subsequent changes to the cell metabolism, based on this simulated example.

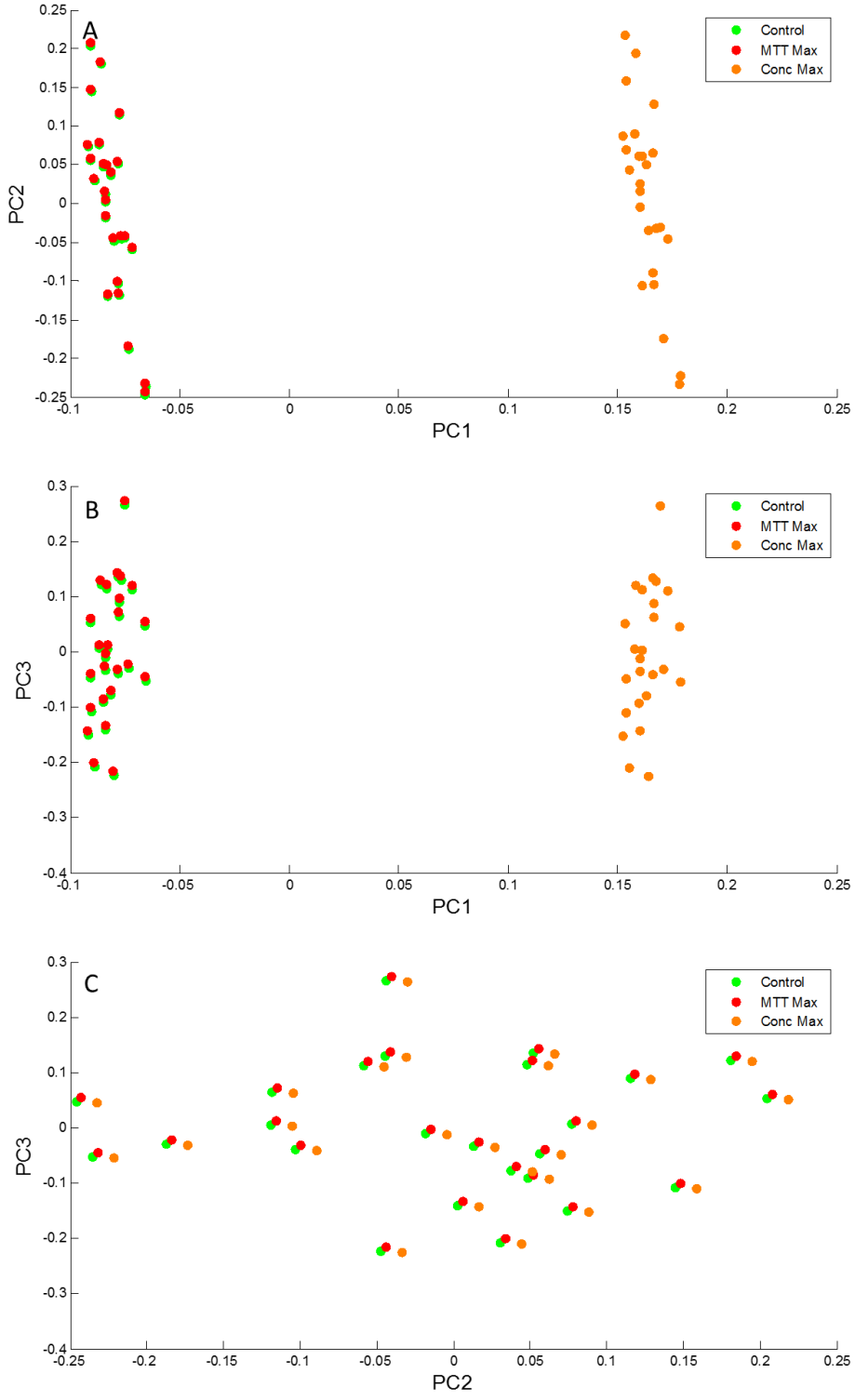


Figure 5.8. PCA of Dataset 2 (A) scatter plot of PC1 vs. PC2 (B) scatter plot showing PC1 vs PC3 (C) scatter plot showing PC2 vs. PC3. The variances describe by PC 1, 2 and 3 are 93.94%, 2.22% and 1.54% respectively for standard PCA. The loadings corresponding to the scatter plots are shown in figure 5.9.

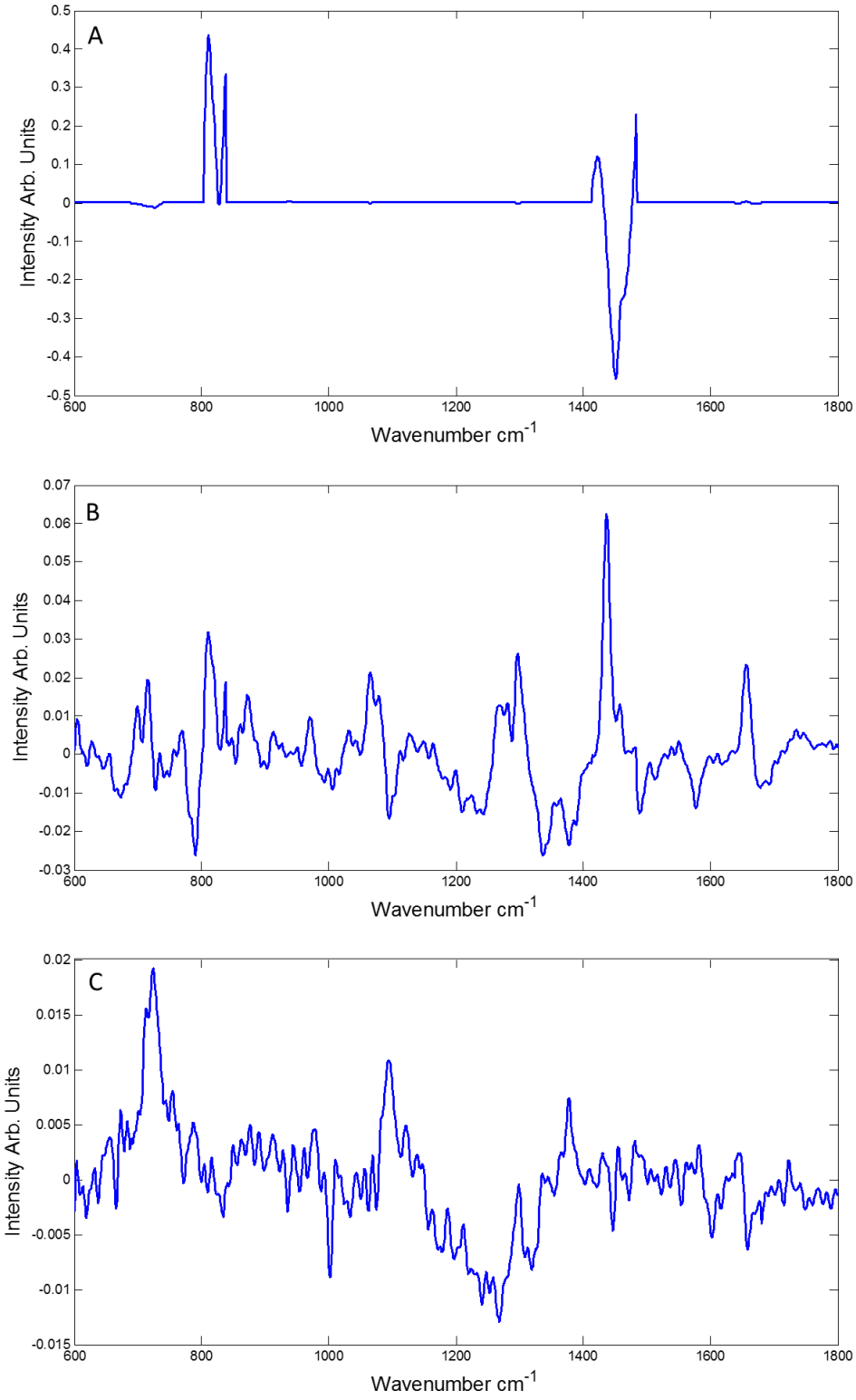


Figure 5.9. Loadings corresponding to (A) PC1, (B) PC2 and (C) PC3 for Seeded PCA on dataset 1. With PC 1, 2 and 3 accounting for 93.94%, 2.22% and 1.54% of the variance respectively.

5.4.5. Seeded PCA Dataset 2

In the case where a spectral profile of a minority variant is known, as in the case of the MTT construct added to the control data, SePCA can be employed, as demonstrated in figure 5.6. To further illustrate the concept for two independent variations, a single spectrum of Construct 2, multiplied by a factor of 10^4 was introduced into Dataset 2, as an additional, independent spectrum. The optimised value of 10^4 was chosen as it allows for the majority of the spectral variance to be described by the first PC (figure 5.5). Thus, Dataset 2 now consists of the control spectra (25), the control + $WMax_{MTT}$ spectra (25), the control + $WMax_{Conc}$ spectra (25), and O_{SW}xC (1), where O_{SW} is the optimised weighting for the addition of the spectral construct.

As shown in figure 5.10, this allows for almost complete separation of each spectral group, control, (Control + $WMax_{MTT}$) and (Control + $WMax_{Conc}$) in Dataset 2. The majority of the variance is described by the first PC; 99.99%, PC2 and PC3 accounting for 0.003% and 0.000079% variance respectively, at this seeded weighting. As a consequence of seeding for the MTT spectral changes, partition of the data is now observed between Control and (Control + $WMax_{MTT}$), according to PC1 (Figure 5.11 A), the loading of which is dominated by the MTT construct of Figure 5.2B, as shown in Figure 5.11A. PC2 is dominated by the loadings of the spectral features of Construct 1, which shows separation along PC2, which differentiates Control and (Control + $WMax_{MTT}$) from (Control + $WMax_{Conc}$). No separation is evident according to PC3, the loading describing background noise.

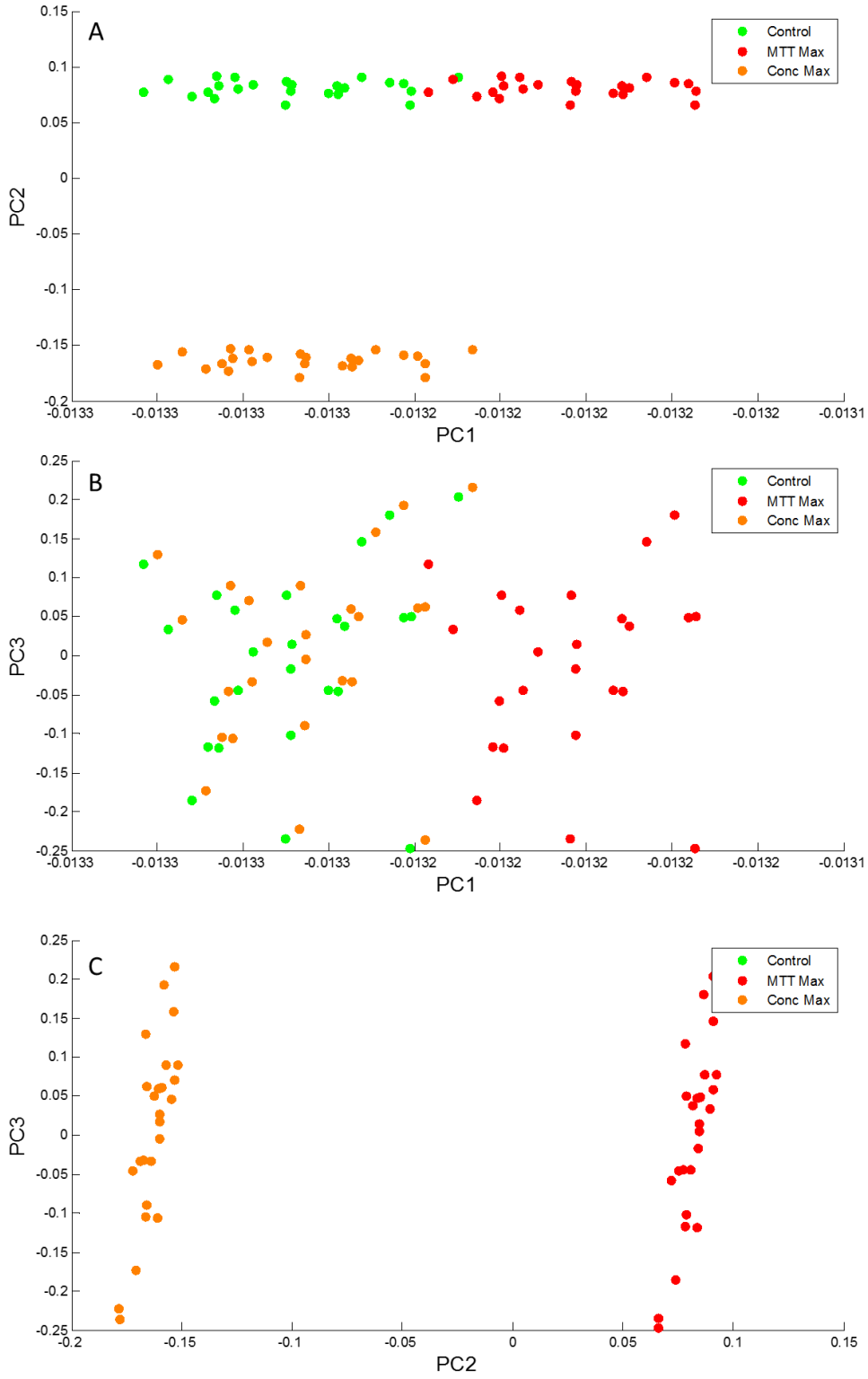


Figure 5.10. Seeded PCA of Dataset 2 (A) scatter plot of PC1 vs. PC2 (B) scatter plot showing PC1 vs PC3 (C) scatter plot showing PC2 vs. PC3. The variances described by PC 1, 2 and 3 are respectively 99.997%, 0.0033% and 0.000079% for seeded PCA.

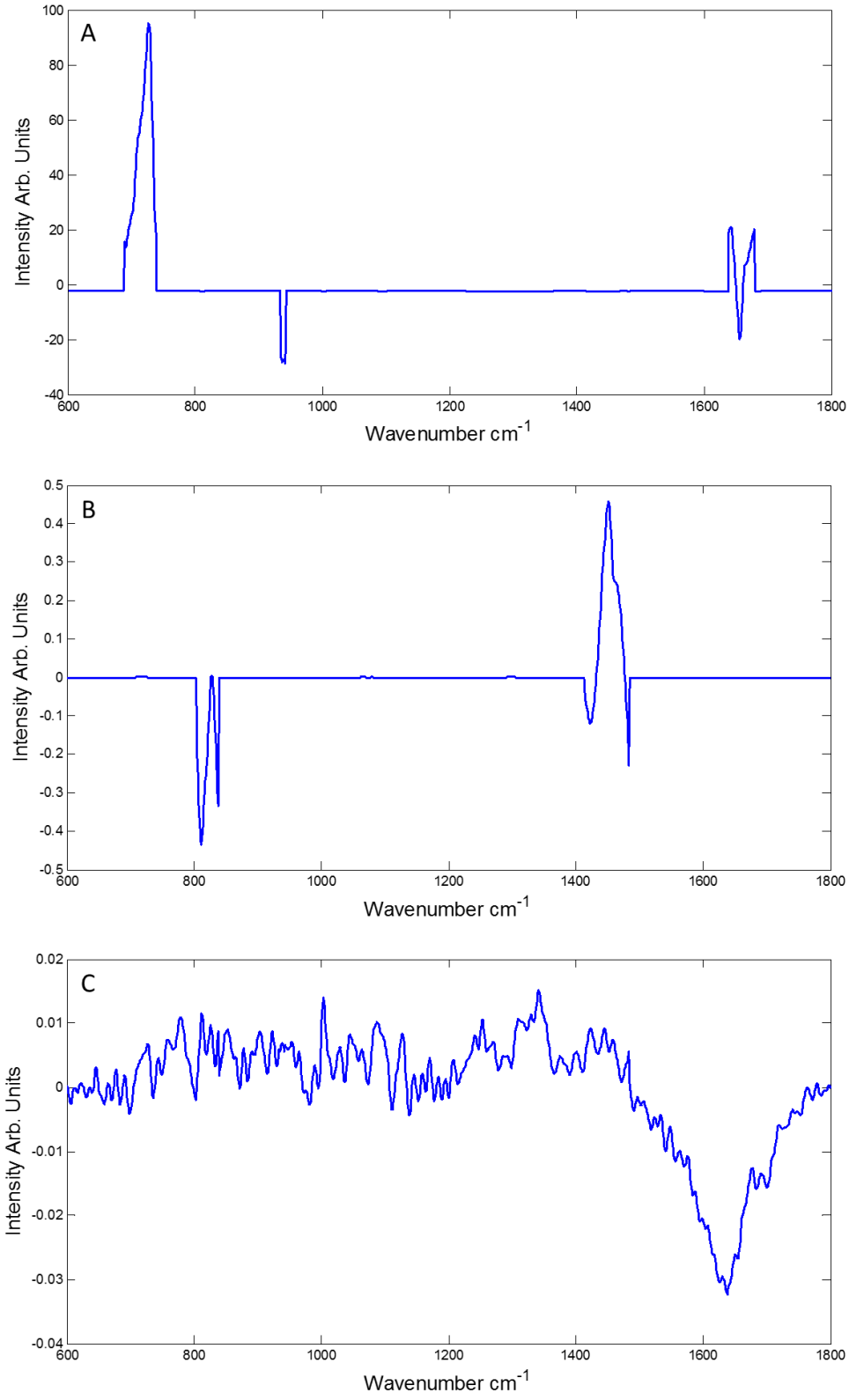


Figure 5.11. Loadings corresponding to (A) PC1, (B) PC2 and (C) PC3 for Seeded PCA on dataset 1. With PC 1, 2 and 3 accounting for 99.997%, 0.0033% and 0.000079% of the variance respectively.

5.4.6. PCA Dataset 3

Standard PCA of dataset 3 is shown in figure 5.12. This dataset consists of systematic changes in both MTT and concentration introduced simultaneously, according to table 5.1. As per previous examples in dataset 1 and 2, construct 2, the MTT experimental changes are not evident in the first three principal component loadings, with the respective variances of 90.99% 7.87%, 0.42% for PC 1, 2 and 3. However, the concentration spectral construct is dominant in figure 5.10 B and is primarily responsible for the separation shown along the PC 2 axis, figure 5.9C. Separation is also shown with the first principal component although the introduced spectral features of the constructs are not apparent in PC 1, figure 5.13A-C.

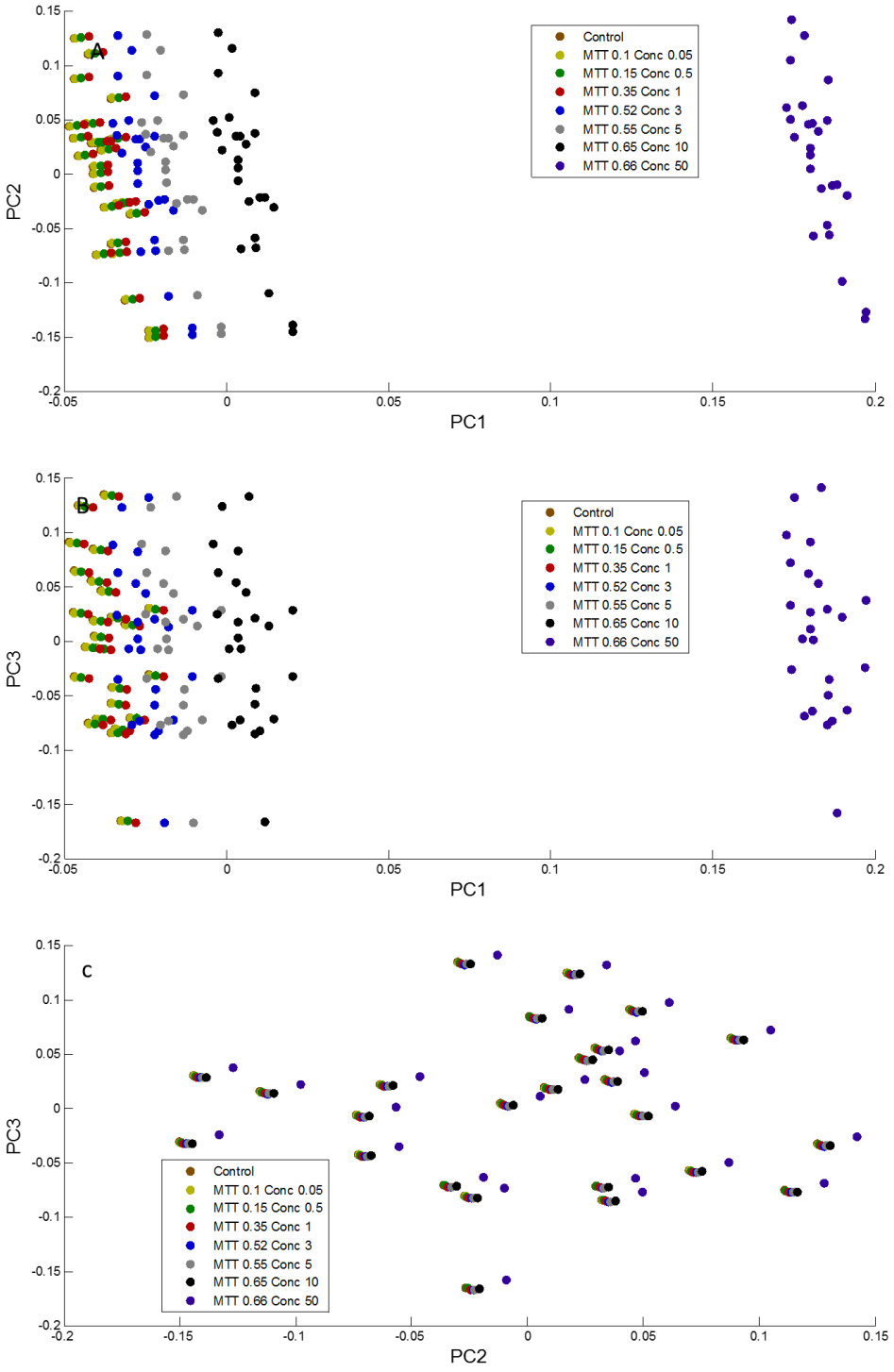


Figure 5.12. PCA of Dataset 3 (A) scatter plot of PC1 vs. PC2 (B) scatter plot showing PC1 vs PC3 (C) scatter plot showing PC2 vs. PC3. The variances described by PC 1, 2 and 3 are respectively 87.82%, 4.51% and 3.13% for PCA.

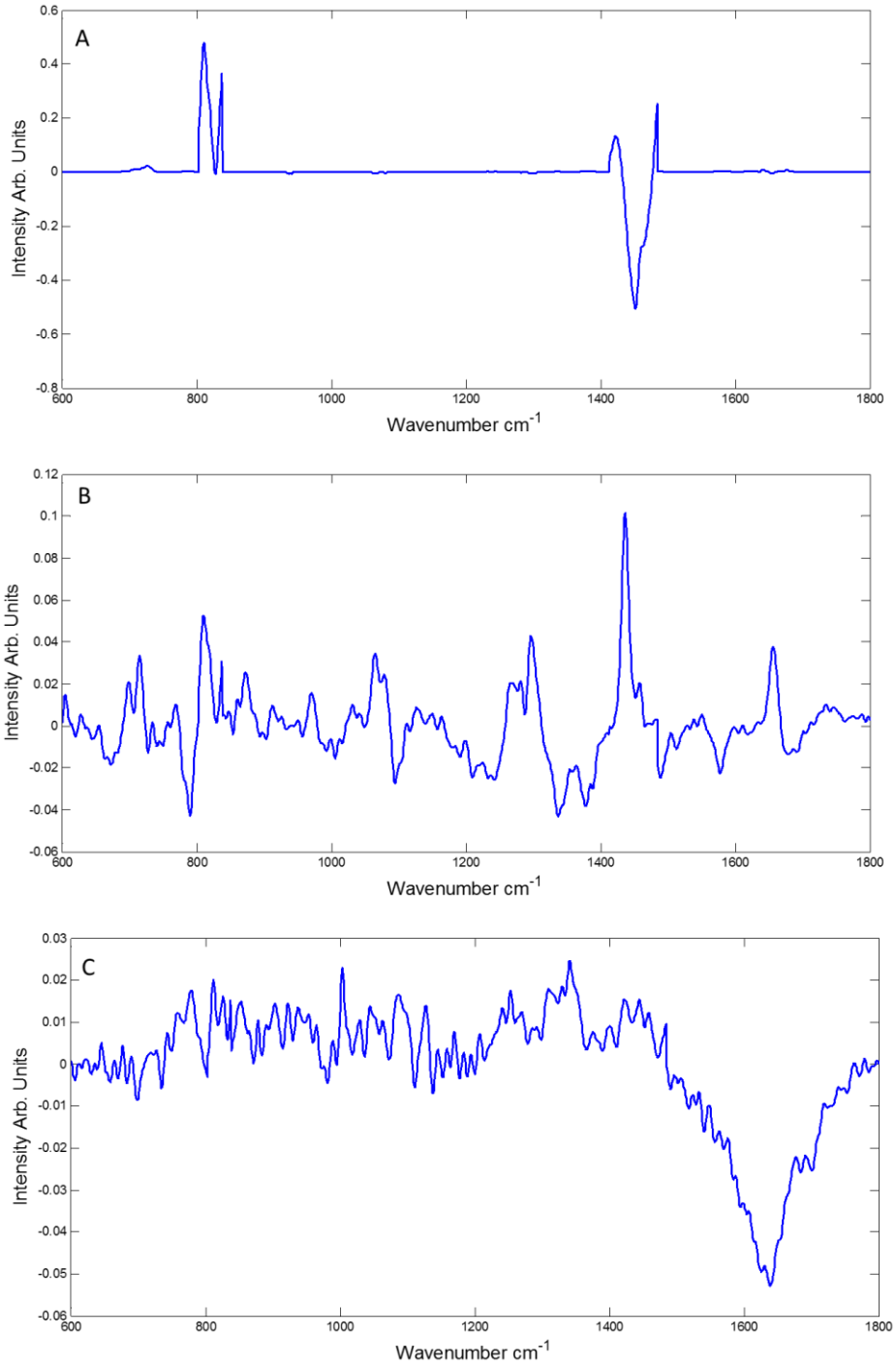


Figure 5.13. Loadings corresponding to (A) PC1, (B) PC2 and (C) PC3 for PCA on dataset 3. With PC 1, 2 and 3 accounting for 87.82%, 4.51% and 3.13% of the

5.4.7. Seeded PCA dataset 3

Seeding the algorithm for Construct 2 allows for both the MTT and Concentration (Construct 1) to be extracted in the first 2 principal components, as shown in figure 5.15 A-C. PC 1 and 2 are responsible for 87.82% and 4.51% of the variance in the data, respectively (figure 5.14 A). PC 3 does not result in any differentiation of the data. Notably, the differentiation of the data for both MTT and Conc variables is continuous, although the degree of separation of each weighting is small, in both cases. As a way of increasing the potential for separation, 1st and 2nd derivative transformations of the data was performed.

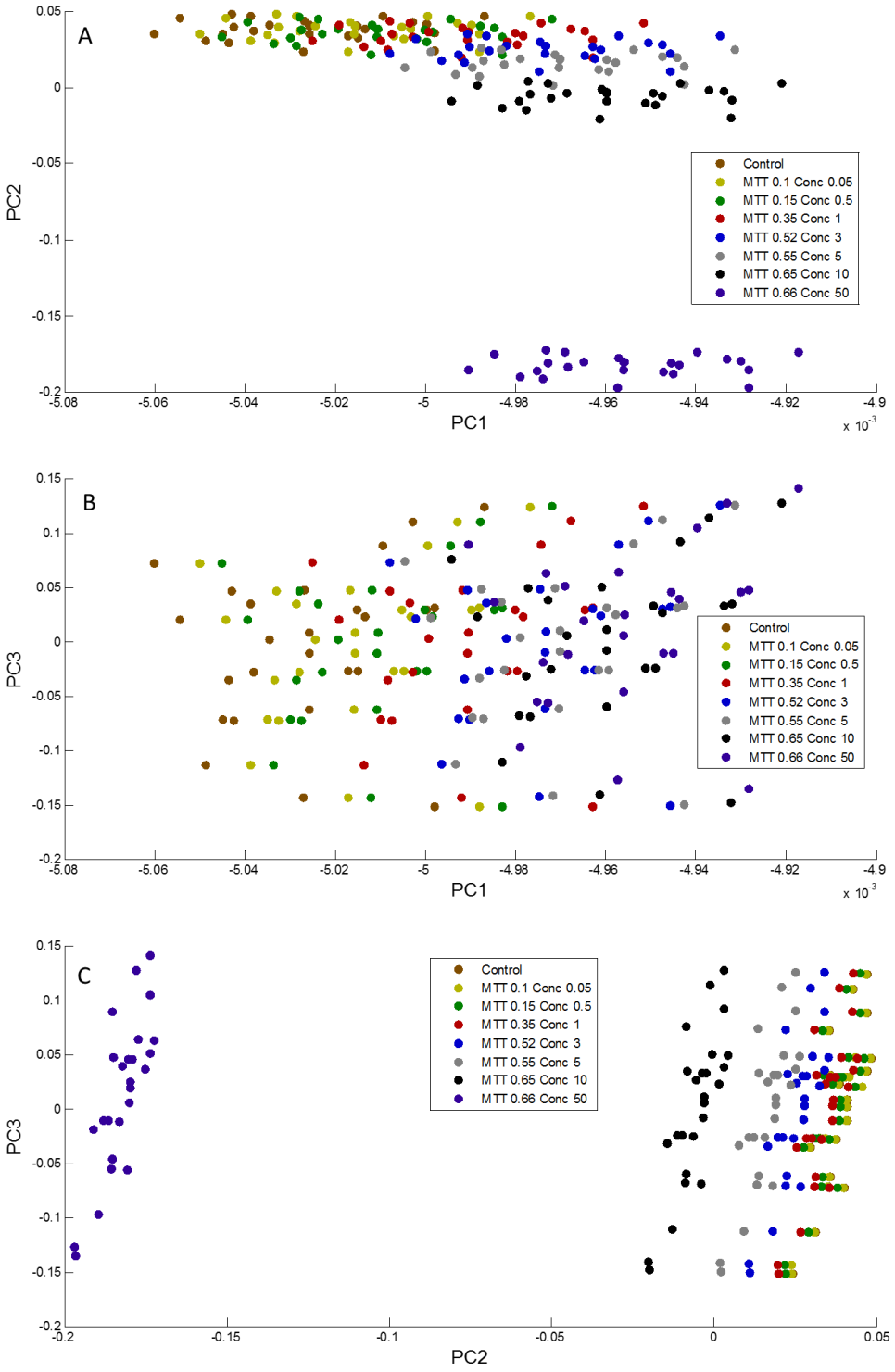


Figure 5.14. Seeded PCA on dataset 3 (A) scatter plot of PC1 vs. PC2 (B) scatter plot showing PC1 vs PC3 (C) scatter plot showing PC2 vs. PC3. The variances described by PC 1, 2 and 3 are respectively 99.99%, 0.004% and 0.0002% for seeded PCA.

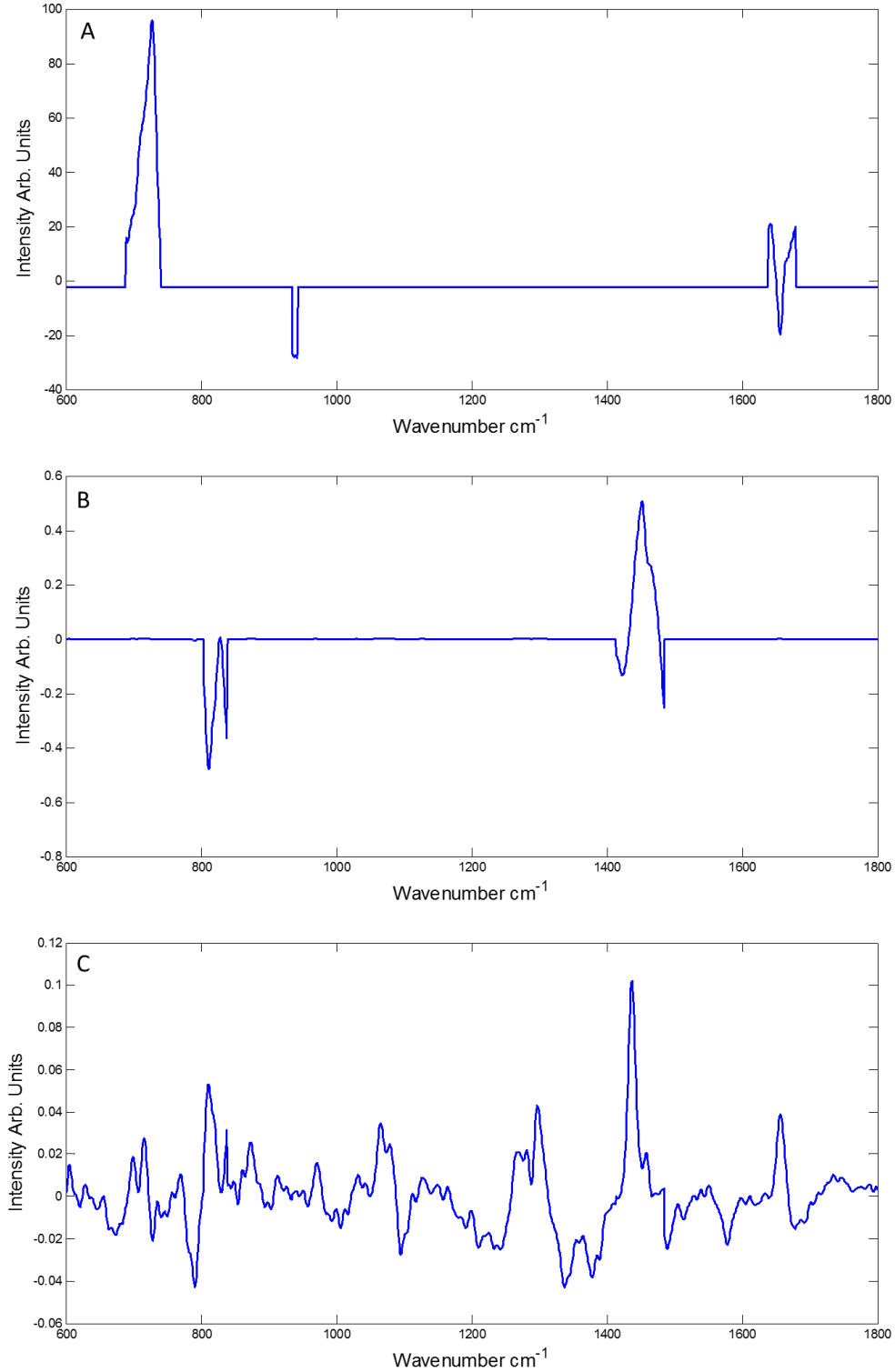


Figure 5.15. Loadings of PCA of Dataset 3 corresponding to (A) PC1, (B) PC2 and (C) PC3. The variances described by PC 1, 2 and 3 are respectively 99.99%, 0.004% and 0.0002% for seeded PCA.

5.4.8. Seeded PCA on 1st derivative spectra

Further improvements are shown for SePCA on 1st derivative spectra from dataset 3 in figure 5.16 A – C, a more evident systematic partition of the data being produced according to PC1 and PC2, but not PC3, with respective variances of 99.99%, 0.0079% and 0.0001%. The loading corresponding to the 1st derivative of Construct 2 is dominant in PC1, which is reflective of the MTT related spectral changes which have been introduced, while PC2 is dominated by the 1st derivative of Construct 1, representative of the Concentration spectra changes introduced (figure 5.17A-C). For standard PCA derivatisation resulted in no separation for dataset 1, and, while for dataset 2 improvements in partition of the data according to variations in Concentration were evident, no features of the MTT construct were extracted.

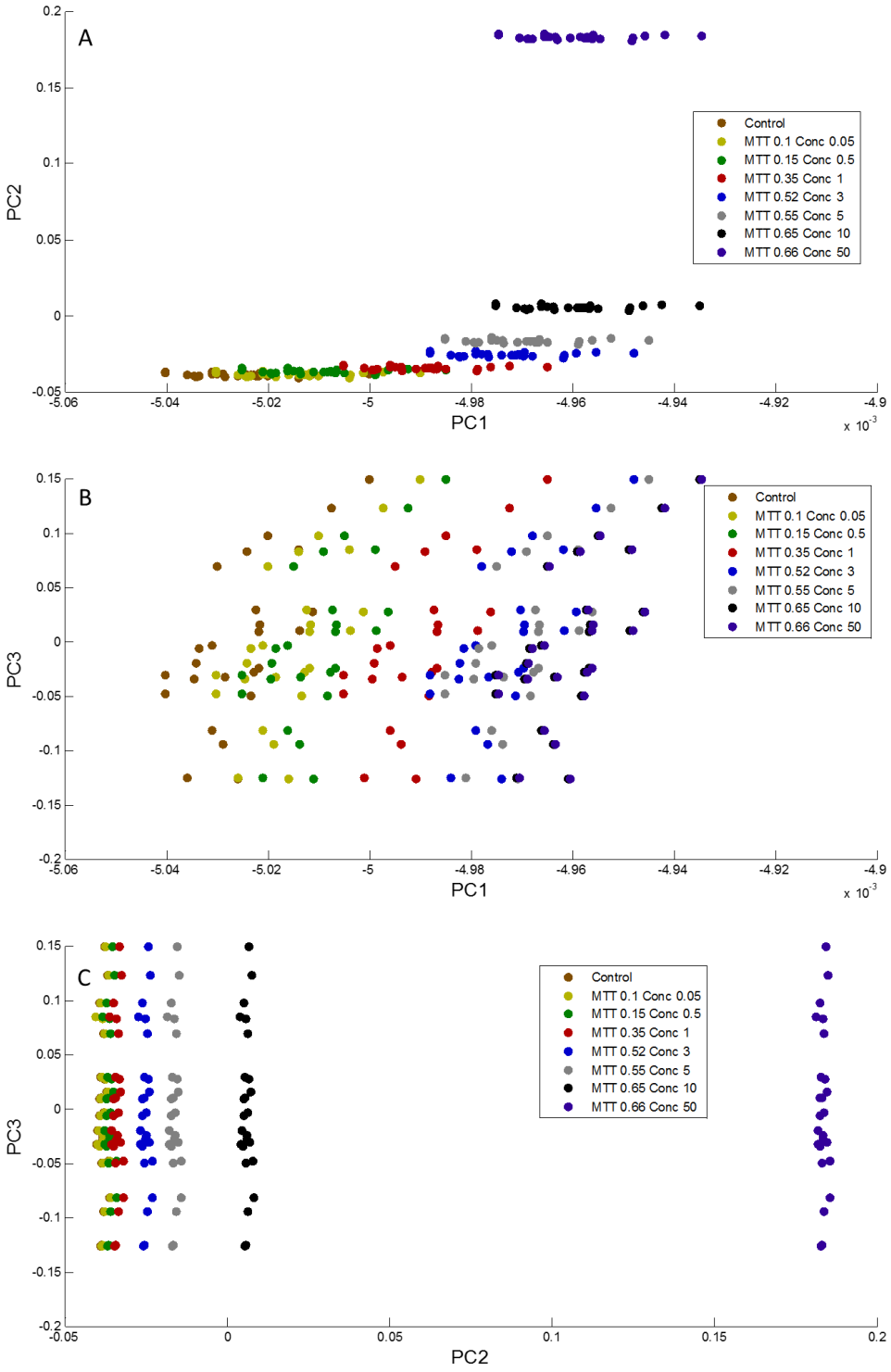


Figure 5.16. Seeded PCA on 1st derivative spectra from dataset 3 (A) scatter plot of PC1 vs.PC2 (B) scatter plot showing PC1 vs PC3 (C) scatter plot showing PC2 vs. PC3. The variances described by PC 1, 2 and 3 are respectively 99.99%, 0.0079% and 0.0001% for seeded PCA.

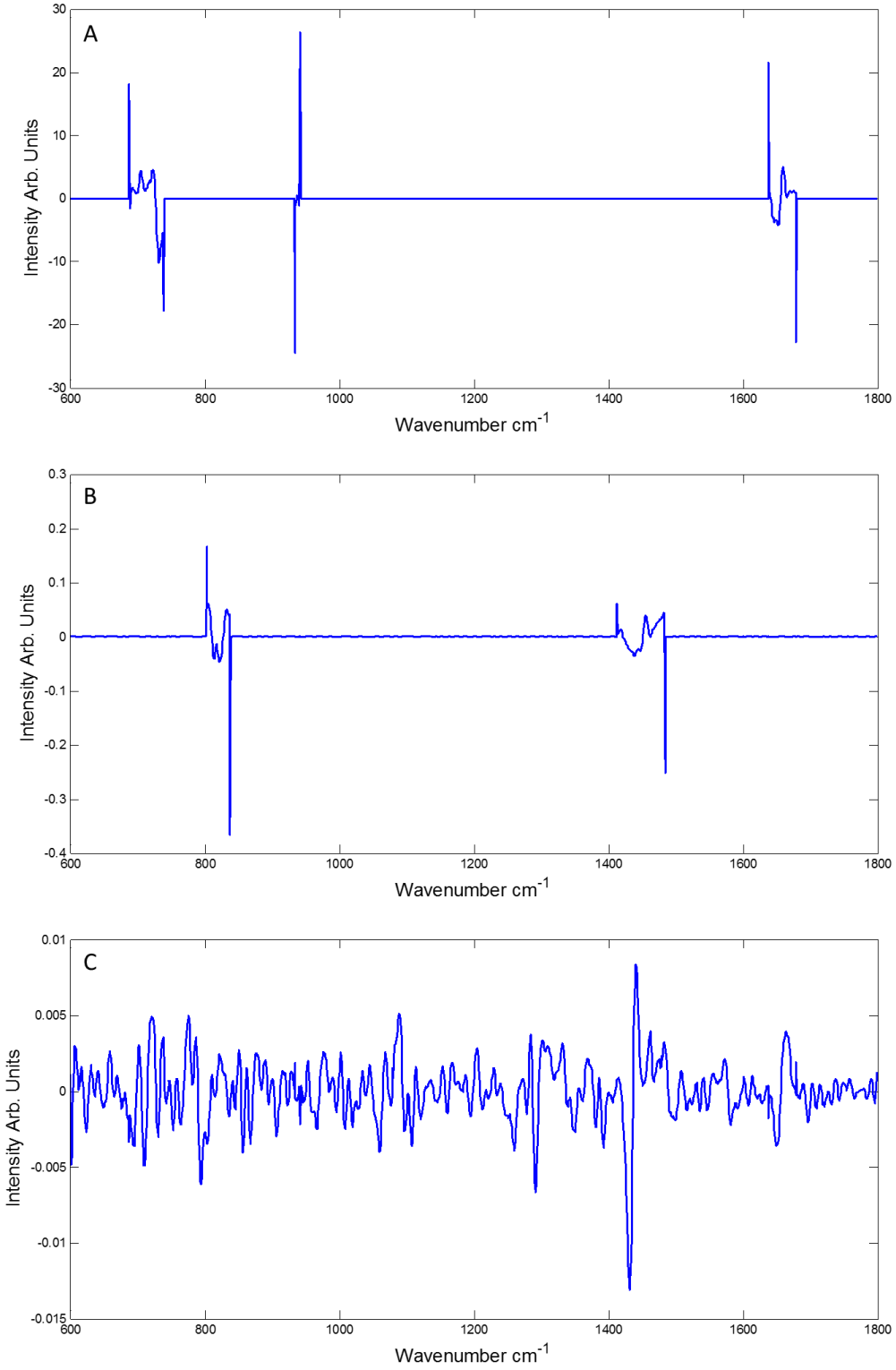


Figure 5.17. Loadings corresponding to (A) PC1, (B) PC2 and (C) PC3 for PCA on dataset 3. With PC 1, 2 and 3 accounting for 99.99%, 0.0079% and 0.0001% of the variance respectively.

5.4.9 Seeded PCA on 2nd derivative spectra from Dataset 3

Considering the improvements in separation achieved following first derivatization of Dataset 3, it was subjected to second derivatization, to explore whether this resulted in a further improvement in the separation of the spectral data. For standard PCA this resulted in no separation for dataset 1, and, while for dataset 2 improvements in partition of the data according to variations in Concentration were evident, no features of the MTT construct were extracted.

Figure 5.18 shows the results following seeding of second derivative of Dataset 3 with the MTT spectral construct, in the same manner as previous analysis. An increased in partition of the data along both PC1 and PC2 is evident and both loadings show pure second derivative loadings, from construct 1 and 2 respectively, figure 5.19 A and B. This enhancement shows the benefits of using second derivative spectra and extracting the correct trend from the spectral data.

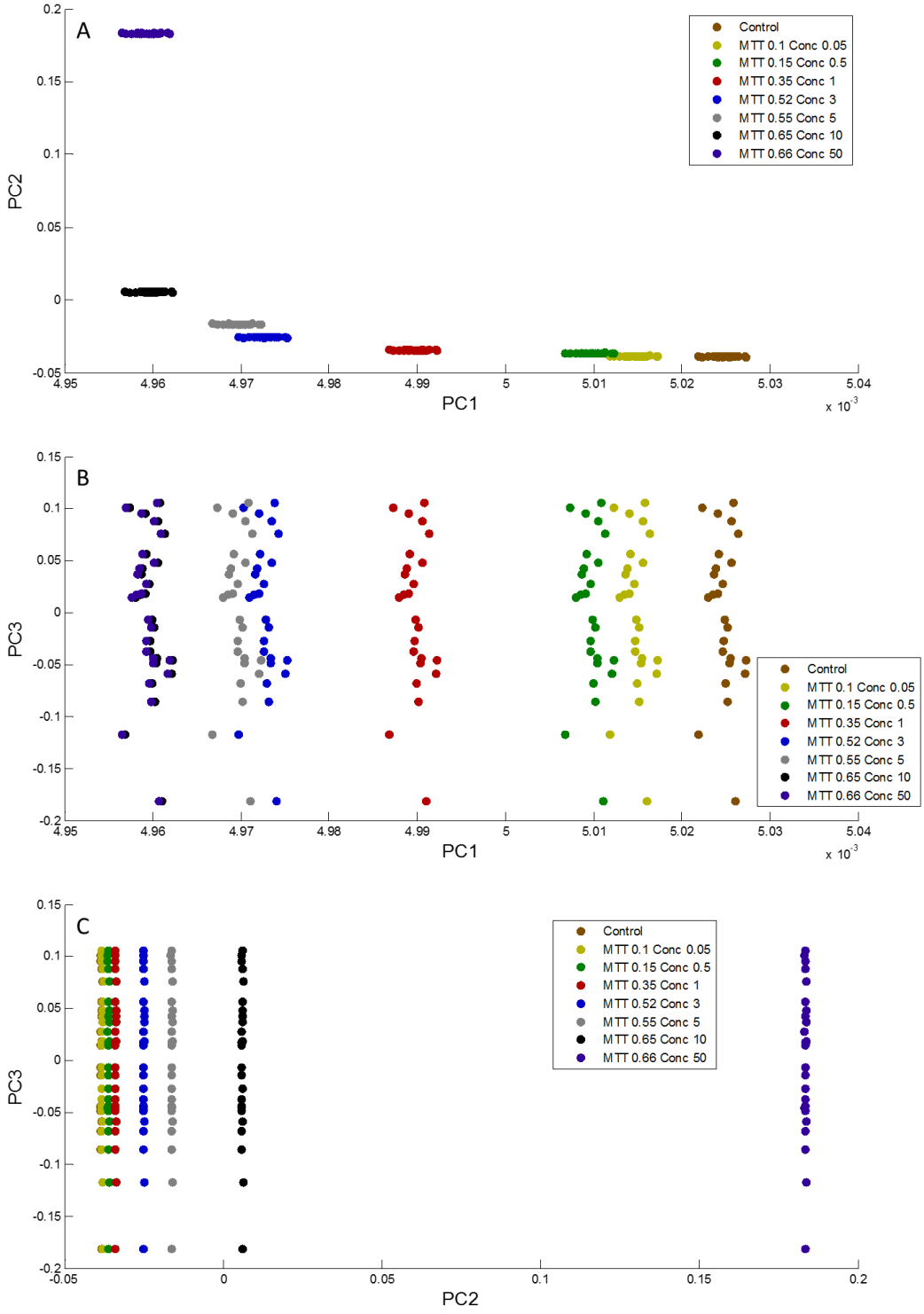


Figure 5.18. Seeded PCA on 2nd derivative spectra from dataset 3 (A) scatter plot of PC1 vs. PC2 (B) scatter plot showing PC1 vs PC3 (C) scatter plot showing PC2 vs. PC3. The variances described by PC 1, 2 and 3 are 99.99%, 0.0087% and 0.000014% for seeded PCA.

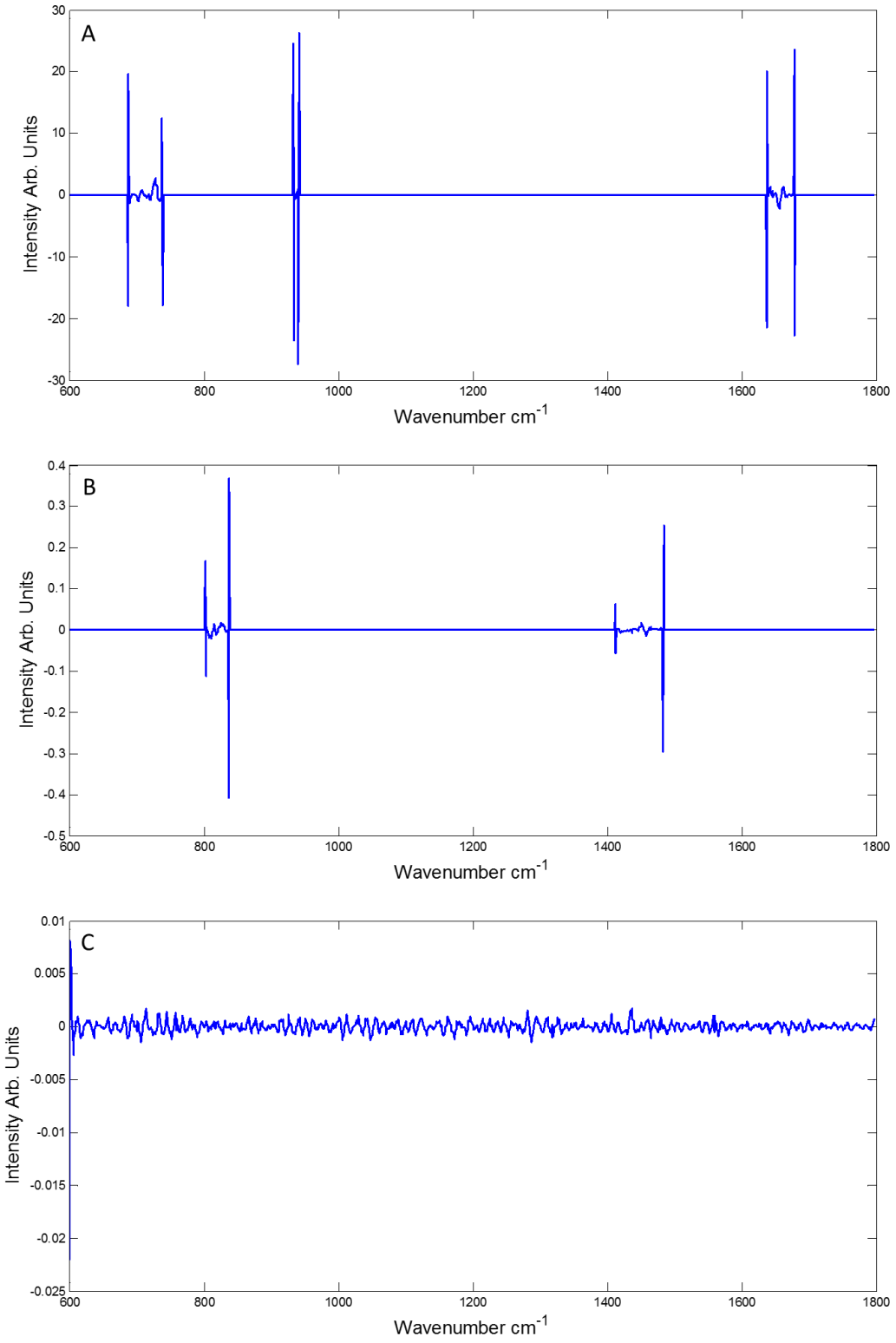


Figure 5.19. Loadings corresponding to (A) PC1, (B) PC2 and (C) PC3 for seeded PCA on 2nd derivative spectra from dataset 3. With PC 1, 2 and 3 accounting for 99.99%, 0.0079% and 0.0001% of the variance respectively.

5.5 Discussion

Validation of multivariate statistical protocols used in vibrational spectroscopy is essential to ensure that the spectral analysis is reliable and accurate for standardised and routine usage. In the development of novel *in-vitro* screening tools for both nano and pharmacological screening, it is imperative that sample preparation protocols, instrumental reliability and multivariate routines are as reliable and accurate as possible, to ensure a smooth transition from lab bench to clinical and company settings.

Focusing on the multivariate statistical analysis, previously published work by Bonnier and Byrne, 2012[27], aimed to elucidate the use and interpretability of PCA in vibrational spectral applications, using both real and simulated data. Similarly, validation and development of the RMie-EMSC algorithm was done using simulated and real data in the work of Bassan et al 2010[31]. These selected examples show the applicability of simulations in validation of multivariate analysis as well as spectral prepossessing in a biomedical vibrational spectroscopic context.

The application of PLSR to independently extract information concerning the direct chemical interaction of chemotherapeutic agents with cells (Construct 1), and the subsequent physiological response of the cells using Raman spectroscopy in parallel with conventional *in-vitro* cytotoxicity assays (Construct 2) was validated using a similar set of simulated spectral datasets by Keating et al, 2015[25]. The current work explores the applicability of PCA to the same simulated spectral dataset, and its ability to extract the systematic and continuously variable spectral perturbations introduced. Limitations of PCA of

the data was shown in figure 5.3 A, whereby the algorithm was unable to extract the desired spectral features from the dataset, as the magnitude of the perturbation was less than the intrinsic variability of the cellular data. A successful partition of the data is shown to be possible when the algorithm is seeded with the known spectral variation, as demonstrated in figure 5.5 A and B. In Dataset 3, which is continuously perturbed by the addition of weighted contributions of the two spectral constructs, seeding the dataset with the minority perturbation enables the continuous differentiation of the data, and extraction of both independent spectral perturbations. Further improvements in separation are shown using 1^{st} and 2^{nd} derivative spectral data for the seeded datasets such that, in the case of the SePCA of Dataset 3, the PCA scatter plot shown in figure 5.20A reproduces somewhat the experimental dose dependent toxicity study of Nawaz et al, 2010. This has implications for in-vitro spectral screening platforms as it shows that the correct trend in simulation can be extracted for the data once the correct features are described to the algorithm i.e. a seeded approach. This is a positive step towards a multivariate dose response curve.

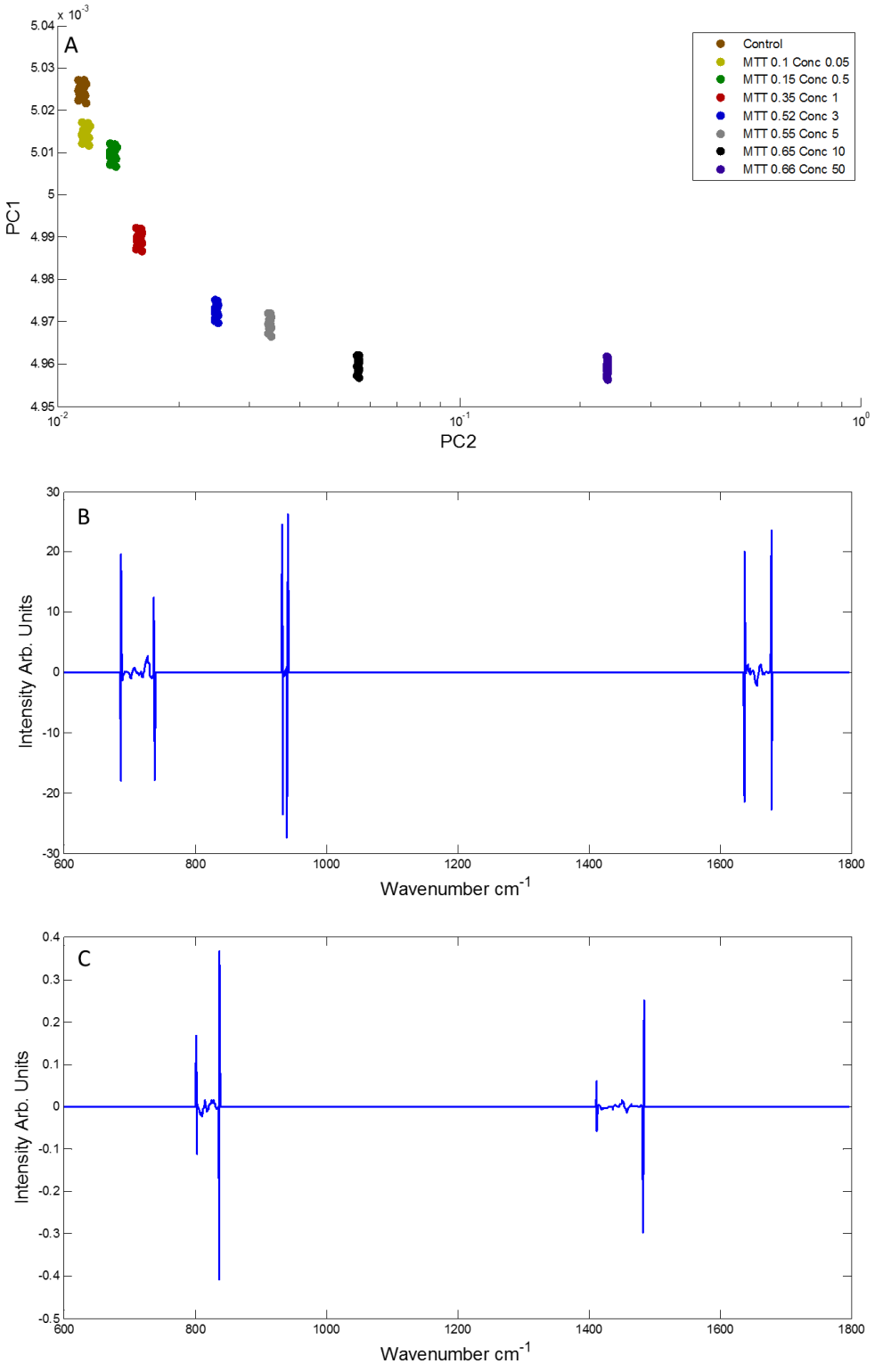


Figure 5.20. Seeded PCA on 2nd derivative spectra from dataset 3 (A) scatter plot of PC2 vs. log PC1 (B) loading for PC1 (C) loading for PC2. The variances described by PC 1 and 2 are 99.99% and 0.0087% respectively for seeded PCA. A constant of 0.05 Arb. Units has been added to PC 2, to allow for log scaling of the data.

The improvements shown have possible ramifications for both diagnostics and *in-vitro* screening. Notably, however, in comparison to the PLSR approach of Keating et al, 2015, the method is supervised, in the context that it requires some prior knowledge of the spectral changes in the data set. In terms of Construct 1, this could be facilitated by a library of spectral signatures of, for example, DNA major and minor groove binders and intercalators, allowing a rapid screening of mechanisms of action of novel chemotherapeutic agents. In a similar fashion, spectral signatures could be established to represent Adverse Outcome Pathways (AOPs), an approach to representation of toxicology recently endorsed by the OECD[32]. In this approach, while the chemical binding of the agent to the receptors represents the Molecular Initiator Event (MIE), cascade of events leading to, for example apoptosis or necrosis constitute the AOP, which could be represented by distinct spectral signatures.

For diagnostic applications such as classification e.g. using support vector machines (SVM) or linear discriminant analysis (LDA), in which PCA coefficients are input to the algorithms, seeding in combination with 1^{st} and 2^{nd} derivative spectra may provide improvements in dividing the data for training and thus, improvements in the diagnostic classification accuracy if the correct variable features can be identified across the patient data.

The nature of the continuously varying spectral changes is also relevant for the interpretation of experimental changes. In this instance (Dataset 3), the changes are continuous and linearly increasing across the entire dataset. However, in experimental data, the changes may not be present in a continuous or linear fashion, or across the entire sampled range. If, as in many instances, the loadings contain an ensemble of spectral features, multiple trends may be responsible for

the pattern of separation in the data. By seeding with the correct peaks the pattern of partitioning in the data can be more accurately identified and adjusted based on the correct spectral changes in the data.

5.6 Conclusions

This study demonstrates an analytical methodology, seeded PCA, which increases the potential of the PCA algorithm to separate spectrally distinct data, particularly in the case where continuous but minor variations are present over a dataset range. The use of 1^{st} and 2^{nd} derivatisation of the dataset is demonstrated to further enhance the differentiation potential of the algorithm. This has important ramifications for improving separation of spectra, with a particular emphasis on biomedical spectroscopy, be that in spectral diagnostics i.e. classification protocols and/or *in-vitro* screening of drugs and nano-materials. The study also demonstrates the benefits of analysis of simulated datasets in the development and validation of novel multivariate analysis algorithms.

5.7 Acknowledgement

This research was supported by the Integrated NanoScience Platform, Ireland (INSPIRE), funded under the Higher Education Authority PRTLI (Programme for Research in Third Level Institutions) Cycle 5, co-funded by the Irish Government and the European Union Structural fund, and Science Foundation Ireland (08/PI/11).

5.8 References

1 H. J. Byrne, M. Baranska, G. J. Puppels, N. Stone, Bayden Wood, K. M. Gough, P. Lasch, P. Heraud, josep Sulé-Suso and G. D. Sockalingum, *Analyst*, 2015, **140**, 2066 – 2073.

2 K. Kong, C. J. Rowlands, S. Varma, W. Perkins, I. H. Leach, A. A. Koloydenko, H. C. Williams and I. Notingher, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 15189–94.

3 K. Kochan, K. M. Marzec, K. Chruszcz-Lipska, A. Jasztal, E. Maslak, H. Musiolik, S. Chłopicki and M. Baranska, *Analyst*, 2013, **138**, 3885–3890.

4 K. Kochan, K. M. Marzec, E. Maslak, S. Chlopicki and M. Baranska, *Analyst*, 2015, 140, 2074–2079.

5 K. Kochan, E. Maslak, C. Krafft, R. Kostogrys, S. Chlopicki and M. Baranska, *J. Biophotonics*, 2015, **8**, 597–609.

6 I. Taleb, G. Thiéfin, C. Gobinet, V. Untereiner, B. Bernard-Chabert, A. Heurgué, C. Truntzer, P. Hillon, M. Manfait, P. Ducoroy and G. D. Sockalingum, *Analyst*, 2013, **138**, 4006–14.

7 M. Xinying, M. Kon, A. Ergin, S. Remiszewski, A. Akalin, C. M. Thompson and M. Diem, *Analyst*, 2015, **140**, 2449–2464.

8 N. Kröger-Lui, N. Gretz, K. Haase, B. Kränzlin, S. Neudecker, A. Pucci, A. Regenscheit, A. Schönhals and W. Petrich, *Analyst*, 2015, **140**, 2086–92.

9 K. Kong, C. Kendall, N. Stone and I. Notingher, *Adv. Drug Deliv. Rev.*, 2015, **89**, 121–134.

10 M. Kozicki, D. J. Creek, A. Sexton, B. J. Morahan, A. Wesełucha-Birczyńska and B. R. Wood, *Analyst*, 2015, **140**, 2236–2246.

11 R. Vyumvuhore, A. Tfayli, O. Piot, M. Le and N. Guichard, *J. Biophotonics*, 2014,19(11) 1–34.

12 F. Bonnier, F. Petitjean, M. J. Baker and H. J. Byrne, *J. Biophotonics*, 2014, **7**, 167–179.

13 J. Desroches, M. Jermyn, K. Mok, C. Lemieux-Leduc, J. Mercier, K. St-Arnaud, K. Urmey, M.-C. Guiot, E. Marple, K. Petrecca and F. Leblond, *Biomed. Opt. Express*, 2015, **6**, 2380.

14 F. M. Lyng, E. O. Faoláin, J. Conroy, a D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan and H. J. Byrne, *Exp. Mol. Pathol.*, 2007, **82**, 121–9.

15 L. M. Almond, *J. Biomed. Opt.*, 2012, **17**, 081421.

16 Y. W. Wang, S. Kang, A. Khan, P. Q. Bao and J. T. C. Liu, *Biomed. Opt. Express*, 2015, **6**, 3714.

17 M. E. Keating and H. J. Byrne, *Nanomedicine (Lond).*, 2013, **8**, 1335–51.

18 H. Nawaz, F. Bonnier, P. Knief, O. Howe, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2010, **135**, 3070–6.

19 H. Nawaz, F. Bonnier, A. D. Meade, F. M. Lyng and H. J. Byrne, *Analyst*, 2011, **136**, 2450–63.

20 Z. Farhane, F. Bonnier, A. Casey and H. J. Byrne, *Analyst*, 2015, **140**, 4212–4223.

21 E. Efeoglu, M. Keating, J. Mcintyre, A. Casey and H. J. Byrne, *Anal.*

Methods, 2015, **7**, 10000-10017.

22 Krau, D. Petersen, D. Niedieker, I. Fricke, E. Freier, S. F. El-Mashtoly, K. Gerwert and A. Mosig, *Analyst*, 2015, 140(7), 2360-8.

23 H. J. Byrne, P. Knief, M. E. Keating and F. Bonnier, *Chem. Soc. Rev.*, 2016, **45**, 1865-1878.

24 T. N. Q. Nguyen, P. Jeannesson, A. Groh, D. Guenot and C. Gobinet, *Analyst*, 2015, **140**, 2439 – 2448.

25 M. E. Keating, H. Nawaz, F. Bonnier and H. J. Byrne, *Analyst*, 2015, **140**, 2482–2492.

26 M. E. Keating, F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 5792–802.

27 F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 322–32.

28 M. MiljkoviÄ‡, T. Chernenko and M. Romeo, *Analyst*, 2010, 2002–2013.

29 M. Hedegaard, C. Matthäus, S. Hassing, C. Krafft, M. Diem and J. Popp, *Theor. Chem. Acc.*, 2011, **130**, 1249–1260.

30 F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 322–32.

31 P. Bassan, A. Kohler, H. Martens, J. Lee, H. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke and P. Gardner, *Analyst*, 2010, **135**, 268–277.

32 Organisation for Economic Co-Operation and Development, *ENV/JM/MONO*, 2013, **6**, 1–45.

Chapter 6: Spectral Cross Correlation as a Supervised Approach for the Analysis of Complex Raman Datasets: The Case of Nanoparticles in Biological Cells

The following is a journal publication in which generation of simulated datasets, data pre-processing, writing of the paper, data analysis and development of cross correlation as a data mining tool in Raman spectroscopy were carried out by Mark E. Keating. Hugh J. Byrne, as supervisor, was primarily responsible for editing and refining of the text as well as guidance with the development of the analytical technique. Sample preparation and data acquisition were carried out as described elsewhere in a paper by Dorney et al[1]. Franck Bonnier was involved with data acquisition and guidance with cross correlation. The format is that of the journal publication, but section and figure numbers have been adapted to the format of this thesis.

Keating ME, Bonnier F, Byrne HJ. Spectral cross-correlation as a supervised approach for the analysis of complex Raman datasets: the case of nanoparticles in biological cells. Analyst. 2012 Dec 21;137(24):5792-802. doi: 10.1039/c2an36169h.

6.1 Abstract

Spectral Cross-correlation is introduced as a methodology to identify the presence and subcellular distribution of nanoparticles in cells. Raman microscopy is employed to spectroscopically image biological cells previously exposed to polystyrene nanoparticles, as a model for the study of nano-bio interactions. The limitations of previously deployed strategies of K-means clustering analysis and principal component analysis are discussed and a novel methodology of Spectral Cross Correlation Analysis is introduced and compared with the performance of Classical Least Squares Analysis, in both unsupervised and supervised modes. The previous study demonstrated the feasibility of using Raman spectroscopy to map cells and identify polystyrene nanoparticles in a lipid rich environment, which is suggestive of the membrane rich endoplasmic reticulum. However, short comings in identification of all nanoparticle signatures in the cell using K-means clustering are apparent, as highlighted by principal component analysis of the identified clusters which demonstrates that K-means clustering does not identify all regions where spectral signatures of the nanoparticles are evident. Thus, two more sophisticated analytical approaches to the extraction of the nanoparticle signatures from the Raman spectral data sets, namely classical least squares analysis and cross correlation analysis, were employed and are demonstrated to improve the identification of spectroscopic signatures characteristic of polystyrene nanoparticles in a cellular environment. Additionally, to investigate the local biochemical environment in which the nanoparticles are trafficked, a pure spectrum of 3-sn-phosphatidyl ethanolamine was cross correlated against the Raman data set, further suggesting the particles are indeed localized in a lipid rich

environment. Furthermore, to demonstrate the robustness and versatility of the analysis method, a spectrum of pure RNA was used to demonstrate that a differentiation could be made between DNA of the nucleus and RNA of the nucleolus using the supervised spectral cross-correlation technique.

6.2 Introduction

Nanotechnology is set to become the first trillion dollar industry in history, with predicted benefits which span a wide range of fields, including applications in site specific delivery of drugs in humans, to antimicrobial paint coatings and textile finishing, to advances in the electronics industry[2–7]. However, there are caveats associated with deploying these nanotechnologies which must be addressed before true realistic applications can be widely accepted and adopted as the norm. It is widely known that nanomaterials, more specifically nanoparticles, possess a range of unique characteristics which in some ways dictate their usefulness and applicability in fields such as medical science. Properties such as increased surface to mass ratio result in an increased reactivity and associated novel optical properties result in new possibilities in diagnostic and theranostic imaging and delivery[8,9], while novel semi-conductor properties are applicable to the electronics industry[10]. However, these properties also potentially have negative implications, most importantly in terms of the potential impact of nanoparticle exposure on human health and the environment. Nanoparticles have been demonstrated to be taken up by cells *in-vitro* and to elicit a toxic response while many reports exist of adverse toxic effects *in-vivo*[11–16].

One of the challenges facing the nanotoxicology community is the detection and monitoring of the interaction mechanisms of nanoparticles in

cells[17,18]. Currently, fluorescent microscopy is the most widely used and accessible method to study nanoparticle uptake and trafficking[19–24]. Necessarily, however, it relies on the use of inherently fluorescent or labelled compounds for visualization and monitoring of nanoparticles inside cells. Most nanoparticles are not intrinsically fluorescent, however, and it has been recently demonstrated that fluorescent labels can be labile, and that the observation and distribution of intracellular fluorescence following nanoparticle exposure is not necessarily representative of the presence or distribution of nanoparticles in the cell[25]. While it is also possible to study the dynamics of nanoparticle trafficking using label free optical microscopic techniques such as dark field and differential interference contrast (DIC) microscopy, the techniques are mainly applicable to metal particles such as gold and silver[26] Transmission electron microscopy (TEM) provides an additional method by which nanoparticles can be visualised in a cellular environment[27–29]. The high lateral resolution obtainable with TEM renders it an ideal method for visualising sub cellular organelles and uptake and interaction of nanoparticles. However, significant sample processing (fixing and ultramicrotoming) is required and only particles with sufficient electronic contrast to the cellular environment can be visualised[29,30]

Thus, a label-free technique is required which can ideally unambiguously identify the presence of the nanoparticles in the cells, their sub-cellular location, and their overall effect on the cellular metabolism. Raman spectroscopy is one such method which may provide an alternate to traditional approaches for studying the nanoparticle-biological interface. The technique provides not only a label free method to visualize how the nanoparticle behaves in a biological environment, but offers the potential to identify the local environment and

simultaneously analyse the associated metabolic changes. To do this, one must combine Raman spectroscopy with analytical data mining approaches to extract the signatures associated with the nanoparticles but also to probe the environment the particles are localized in, and to correlate the exposure and subcellular interaction mechanisms with the metabolic changes.

Previous studies have indicated the potential of Raman as a label free method for studying biological processes. Examples include novel approaches for cervical cancer diagnostics[31], to investigating the effects following exposure to human papilloma virus (HPV) infection[32], the effects of chemotherapeutic anticancer agents in cells[33,34], live cell analysis[35,36]and the toxic responses to single walled carbon nano-tubes (SWCNT), to name but a few[37].

Surface enhanced Raman scattering (SERS) is also a potential method to study the intracellular dynamics of nanoparticle trafficking and compartmentalisation[38,39]. However, only certain types of nanoparticle, such as gold and silver particles and nanoaggregates have the potential to generate SERS spectra, thus limiting the technique to the study of only a certain type of nanoparticles. Additionally the surface enhancement process and molecular specificity of the technique are not fully understood, which may lead to ambiguity in the understanding of cellular trafficking.

A more recent study indicated the ability of Raman spectroscopy to detect the presence of intracellular polystyrene nanoparticles[1]. Polystyrene was chosen as a model nanoparticle for the study as it is commercially available and regularly employed as a standard in nanotoxicology (particularly as a positive control in its aminated form). Furthermore, the conjugated styrene ring makes it a relatively strong Raman scatterer. However, while the identification is somewhat straight

forward, the presence of overlapping peaks in both the polystyrene and cellular spectra (e.g. both cellular and polystyrene spectra exhibit a strong symmetric ring breathing peak at ~1004cm^{-1}) presents a challenging system with which to validate the effectivity of the experimental and data analysis techniques. K-means clustering analysis (KMCA) analysis was used to differentiate regions of the cell as well as to identify and localise the nanoparticles. Analysis of the local cellular environment of the detected nanoparticles was performed via a comparison between loadings obtained from principal component analysis (PCA) and pure spectra of lipids and polystyrene nanoparticles. However, when the data was analyzed using PCA, it was noted that the clusters detected using KMCA failed to identify all regions which contained the spectral fingerprint corresponding to polystyrene in a biological environment. Furthermore, the average spectra of the cluster identified by KMCA, while containing features clearly characteristic of polystyrene, also contained spectral features of the neighbouring cellular environment. Analysis of the loading of the principal components provided a clearer differentiation of the nanoparticle contributions from the local cellular environment, but neither unsupervised technique provided an unambiguous localisation of the target species[39].

Other multivariate analytical approaches have also been applied in the field of Raman microspectroscopy of cells. In addition to KMCA, other clustering methods such as Fuzzy C means clustering (FCM) and hierarchal cluster analysis (HCA) have been used to separate the cellular Raman data into clusters and subsequently reshape the data into images[40,41]. However, as highlighted by Headegaard *et al.*, these approaches have their own limitations. In particular boundaries between sub-cellular features can often result in the addition of extra

clusters with mixed spectral signatures. This addition can be overcome by increasing the number of clusters; however, this in turn can result in added complexity to interpretation and inaccuracies in regional separation. Additionally, the reproducibility of these methods can also be questioned as the starting point for the centroid based KMCA and FCM is subjective[40].

PCA and vertex component analysis (VCA) have also been used to separate out distinct regions of the cell. With regards to PCA, separation is based on the variances between the spectra in the data set, the majority of the variance being described by the first three principal components[40]. Thus, the score values can be used to construct a composite image of the cell in which the biochemical contributions of each component are described by the corresponding loadings plot. Unlike KMCA and FCM, PCA identifies quite accurately the boundaries between each feature. However, the images generated suffer from inferior contrast and in some instances interpretation may be difficult as biochemical features may be spread across different loadings.

VCA is another method which has been used for similar analytical purposes. In brief, VCA computes a linear combination of supposed pure component spectra which are termed endmember spectra. As described in Miljkovic *et al.*, the endmember spectra are acquired under the assumption that the most extreme data points in the dataset are representative of pure component spectra[41]. However, it has been pointed out that the endmembers generated are not truly representative of the pure component they describe in the data set and can often contain a mixture of biochemical constituents i.e. DNA and proteins[42]. While this is representative of the true nature of nucleic acids *in-situ*, it could lead to inaccuracies in interpretation.

The work presented here demonstrates the potential of a Spectral Cross Correlation Analysis (SCCA) for the analysis of Raman spectral datasets. The method is applied to the dataset of Dorney *et al.*[1], of polystyrene nanoparticles in A549 lung adenocarcinoma cells, and is thus compared with previous analyses by KMCA and PCA. The performance of SCCA is also compared to that of classical least squares analysis (CLSA), performed both in a supervised and unsupervised manner, which allows for a direct comparison between both approaches. SCCA utilises the spectrum of the target chemical component and cross correlates the spectrum with that of the complete Raman spectral dataset. The quantitative performance is demonstrated using simulated datasets and the potential is demonstrated by mapping the spatial profile of the polystyrene nanoparticles in the cells as well as other biochemical components of the cell, (RNA and lipids).

6.3 Experimental

6.3.1 Sample Preparation for Raman Imaging

A549 Cells were seeded at a density of 4 x 10^4 cells onto calcium fluoride (CaF_2) windows (Crystran Ltd., UK) for confocal Raman imaging. The cells were incubated for 24 hrs in Dulbecco's Modified Eagle's Medium (DMEM F12), supplemented with 10% foetal calf serum (FCS) and 1% L-Glutamine at 37°C, 5% CO_2. Following cell adherence, 2 mLs of medium containing 1x 10^{12} nanoparticles per mL were added to the cells. The cells and nanoparticles were incubated for 24hrs at 37°C and 5% CO_2. Following nanoparticle exposure, the cells were washed in warm PBS three times and fixed for 10mins in 10% buffered formalin. After fixation, the cells were washed to remove any trace of fixative and kept in NaCl solution prior to imaging.

Component spectra used in SCCA were generated as described in Bonnier and Byrne 2012[43]. For polystyrene nanoparticle spectra, nanoparticle suspension was added drop-wise to a CaF_2 window and allowed to air dry prior to Raman acquisition. RNA from baker's yeast (*saccharomyces cerevisiae*) was added to water and subsequently deposited on a CaF_2 window and allowed to air dry. 3-sn-phosphatidyl ethanolamine was dispersed in chloroform and deposited on CaF_2 windows.

6.3.2 Confocal Raman Spectroscopic Imaging

Confocal Raman Spectroscopic Imaging was performed using a Horiba Yobin-Yvon LabRAM HR800 spectrometer with a 785nm, 300mmW diode laser as source and a Peltier cooled 16-bit CCD. A 100X, N.A. 1.2, (LUMplanF1, Olympus) water immersion objective was used for all cellular measurements. The confocal pin hole of the system was set to 100µm, the recommended setting for confocal operation, to allow optical sectioning of the sample. A 300 lines per mm spectroscopic grating, providing a dispersion of ~1.5cm^{-1} per pixel, was used and the system was pre-calibrated to the spectral line at 520.7cm^{-1} of silicon. Using an automated programmable stage, Raman spectra of the cell were acquired with a 0.75µm step size over a 29*39 pixel area which encompassed the nuclear, perinuclear and cytoplasmic regions of the cell.

6.3.3 Data Pre-Processing and Preparation

In order to prepare the data for analysis, a number of steps were taken to ensure the spectra in the map were of a high enough quality to give accurate results. For CLSA, all data pre-processing was carried out using Labspec 5 software which comes as standard on the Raman instrument. Firstly, a background spectrum

which constituted the contribution of the CaF_2 substrate and water in the imaging medium was subtracted from each spectrum in the mapped data set. Following subtraction of the background spectrum, a Savitsky-Golay smoothing filter (5^{th} order, 7 points), available on the software, was used to lightly smooth the data. The data was then baseline corrected using a nodal point baseline correction using the minimum amount of points possible to ensure minimal alteration of the acquired data. Normalization was carried out automatically by the software during CLSA.

Data was prepared in a similar fashion for SCCA. However, the pre-processing was carried out in Matlab (Mathworks,USA) using previously published protocols for data processing[1]. As outlined above, a background spectrum was subtracted from the Raman data set to remove the substrate and immersion medium contributions. A Savitsky-Golay smoothing filter (5^{th} order, 7 points) was applied to the data and a nodal point baseline correction was used to baseline the data using a minimum amount of reference points to do so. Preparation of component spectra for SCCA was done in the same manner for polystyrene, RNA and lipids.

6.3.4 Classical Least Squares Analysis

CLSA was carried out using Labspec 5 software which comes as standard on the Raman spectrometer software. The analysis method is based on a fit of a linear combination of reference component spectra to the spectra contained in the raw spectral map. This is described by Equation 6.1, for the case where three reference component spectra are used. S is the sum of the linear contribution of the

reference components (A, B, C), and x, y, z are the respective weightings or scores necessary for the weighted sum of the reference component spectra to match the raw data.

$$S = [x*A] + [y*B] + [z*C] \qquad \text{Equation 6.1}$$

Using the software, there are two different ways to obtain the reference component spectra. The first way is to obtain a pure spectral reference from a compound or compounds which can then be fitted according to Equation 6.1. The second method uses a factor analysis algorithm to generate the component spectra, the weighted sum of which is compared to the Raman spectral data set. Using the latter of the two methods, Zavaleta et al demonstrated the power of the technique to quantify quantum dot accumulation in an *in-vivo* mouse model and to separate out the different spectral contributions from complex SERS signals in the same data set[44]. In a similar and different way, both approaches to CLSA are explored to extract spectra which contain polystyrene nanoparticles and define other biochemical regions such as the RNA and lipid rich environments. The relative contributions of the different components are defined by the weighting factors (x, y, z….).

6.3.5 Spectral Cross Correlation Analysis

For SCCA, reference spectra from polystyrene, phosphatidyl-ethanolamine and RNA (Figure 6.1 A) were used to screen the Raman spectral data set. All SCCA was carried out using Matlab (Mathworks, USA) using the "crosscorr" function available in the signal processing toolbox. Equation 6.2 describes the cross

correlation between two data series, where C(x) is the correlation function, S(t) is the Raman spectrum in the data set to be tested and A(x+t) is the reference spectrum i.e. polystyrene, lipid or RNA. The function integrates the product of the two data series (spectra) at each point as they are shifted relative to each other along the x axis (wave number). The magnitude of the correlation quantifies the relative contribution of the component spectrum at that point in the cell, and an exact correlation occurs when the spectra are exactly matched (auto-correlation). In this way, it is possible to screen the map or spectra in the map and, based on the cross correlation function, cluster different biochemical regions of the cell based on the relative contributions of the reference spectrum used.

$$C(X) = \sum_{m=-\infty}^{\infty} S(\tau).A(X+\tau) \qquad \text{Equation 6.2}$$

6.3.6 Simulated Data

Simulated data sets were used to test the robustness and sensitivity of both CLSA and SCCA in their ability to detect spectral contributions due to polystyrene, RNA and lipid in a biological environment. To generate the simulated data sets, a cellular spectrum was used as a template to which varied amounts of component spectrum were added. Keeping the cellular spectrum constant, a series of 38 simulated spectra of ratios 1:1 to $1:10^{-4}$, cellular: component Raman spectra for polystyrene, RNA and lipid were generated (Figure 6.1A). An example of the simulated data set for polystyrene is shown in Figure 6.1B, which shows the addition of the first 8 spectral dilutions to the constant cellular spectrum. Using these simulated datasets, it was possible to explore how each data mining approach performs when testing experimental data and thus facilitate accurate interpretation of the data sets.

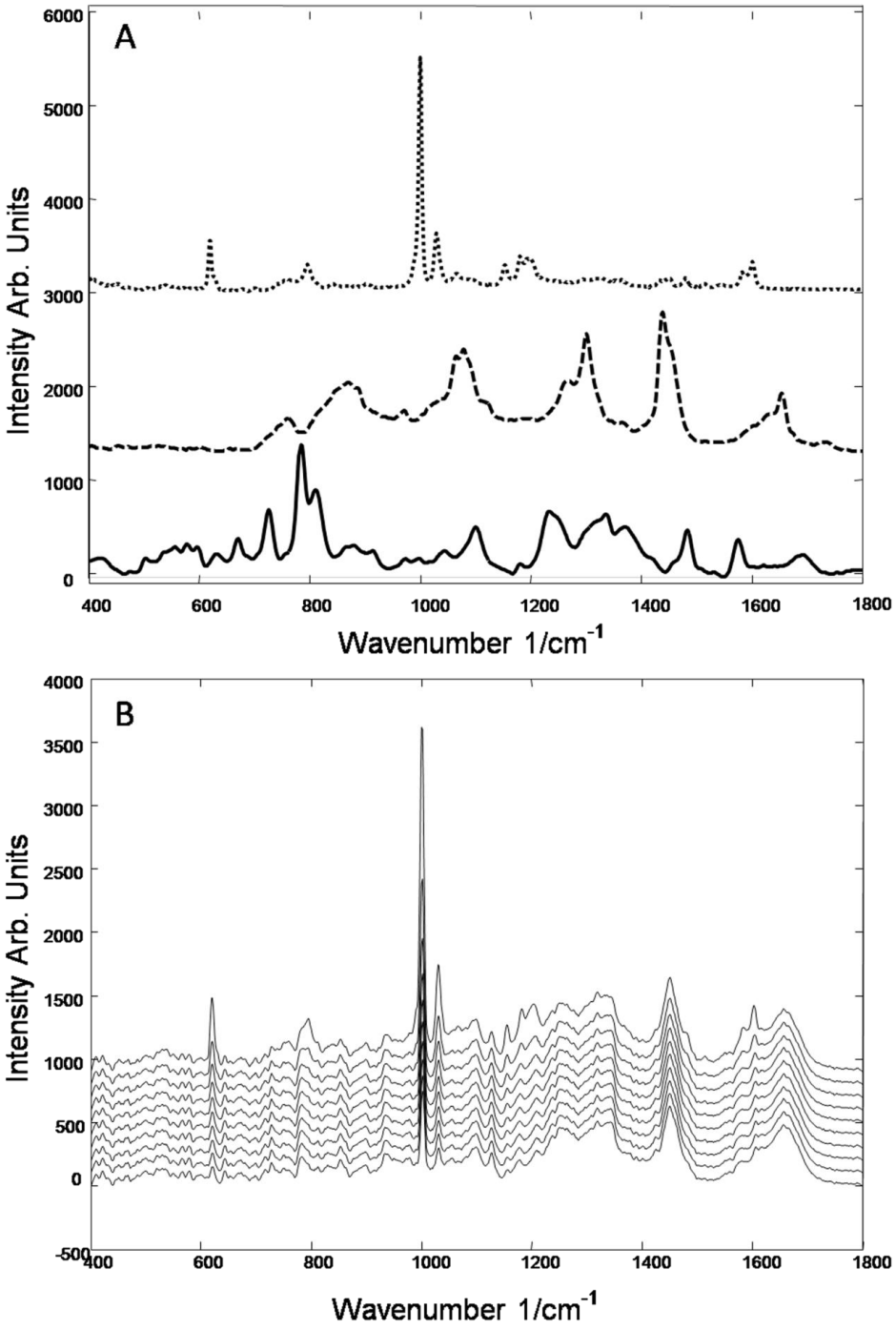


Figure 6.1 (A) Component spectra of nano-polystyrene (dotted line), 3-sn-phosphatidyl ethanolamine (dashed line) and isolated RNA (solid line), offset for clarity. (B) Shows an example of the first eight simulated spectra for polystyrene in cells, offset for clarity. Each spectrum consists of a constant cellular spectrum with a varied concentration of polystyrene added to it, with decreasing

polystyrene concentration from top to bottom. Simulated data sets generated in this way were then analysed by CLSA and SCCA.

6.4 Results

6.4.1 Simulated Data – Unsupervised CLSA

CLSA can be carried out in two different ways, either by generating spectral models using a factor analysis algorithm (unsupervised), or by manually inputting the component spectra (supervised). The data in Figure 6.2 shows the results using the factor analysis generated models for simulated data sets generated based on cellular/polystyrene, RNA and lipid spectra (Figure 6.2 B). In each instance, the score recorded form CLSA for each spectrum is plotted against the component concentration added to the data set. In all cases, the extracted CLSA scores accurately represent the true component ratios over the concentration range, represented by the solid line. The results depart from nonlinearity a cellular:component ratio of ~1:0.1, after which the CLSA weightings no longer accurately reflect the correct component weighting, although the presence of the component can still be identified in ratios as low as 1:0.03.

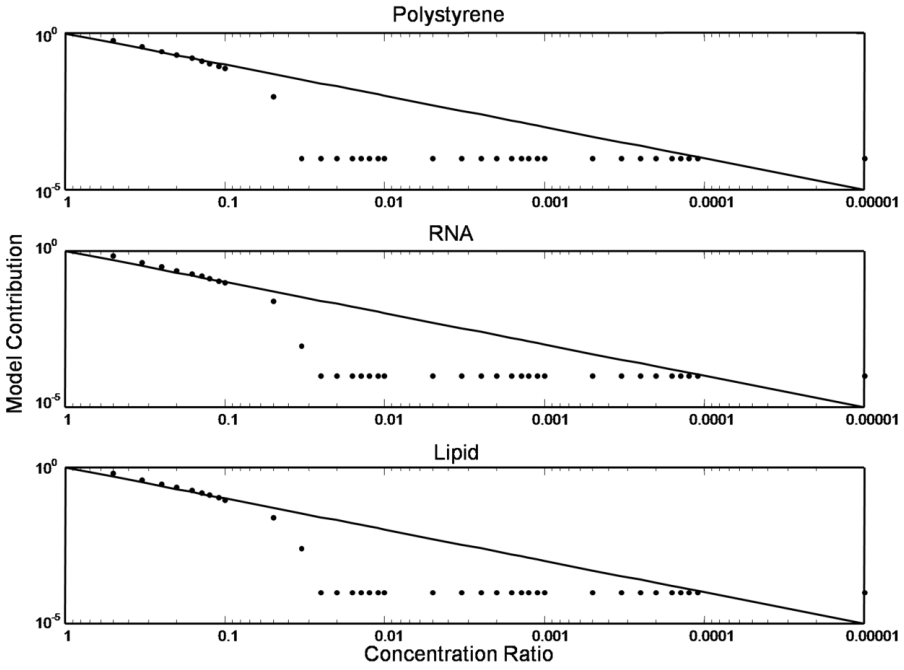


Figure 6.2 CLSA of simulated spectral data sets of nano-polystyrene, RNA and lipid. In each graph, the score from the CLSA is plotted against the concentration of component spectrum added to a constant cellular spectrum (points on each graph). The solid black line represents the ideal response which gives an indication of the quantitative nature of the technique.

6.4.2 Single Cell Data – Unsupervised CLSA

In order to further test the ability of CLSA to identify intracellular polystyrene nanoparticles located inside a single cell Raman map, an initial factor analysis algorithm was applied to the data set to generate 7 model spectra to be used in the CLSA. These model spectra were then used to compute the scores from the Raman data set (Figure 6.3 A). It is then possible to segment the cell into different distributions based on specific spectral differences as shown in Figure 6.3 B. The spectral profile of each model contribution can be visualized individually showing the percentage contribution at each pixel (Figure 6.3 C-F). A more detailed look at the model spectra generated and corresponding cellular distribution can be seen in Figure 6.3 A-G.

The CLSA map shows a different spatial distribution of each model in the Raman spectral data set. Although in all cases, the model spectra show strong contributions of the cellular environment, they are differentiated by contributions from distinct components. Model 1 (Figure 6.4 A) shows characteristic peaks corresponding to those seen in pure polystyrene spectra (see Figure 6.1 A). Therefore, the pixel distribution of model 1 is deemed to show the localisation of the polystyrene nanoparticles, indicating a perinuclear distribution in the cell, consistent with the K-means cluster analysis of Dorney et al[1]. Other models show a different distribution in the cell. Model 6 shows a distribution which visually corresponds to the nucleolus of the cell (Figure 6.4 B), whereas model 3 surrounds the nucleoli and is identified as the nucleus of the cell (Figure 6.4 E). This shows the ability of CLSA to differentiate the biochemical regions of the cell containing RNA and DNA. Other models such as model 4 (Figure 6.4 C) and model7 (Figure 6.4 F) show a distinct distribution surrounding the nucleus, which may correspond to perinuclear organelles such as the endoplasmic reticulum or the Golgi apparatus which are lipid rich regions of the cell.

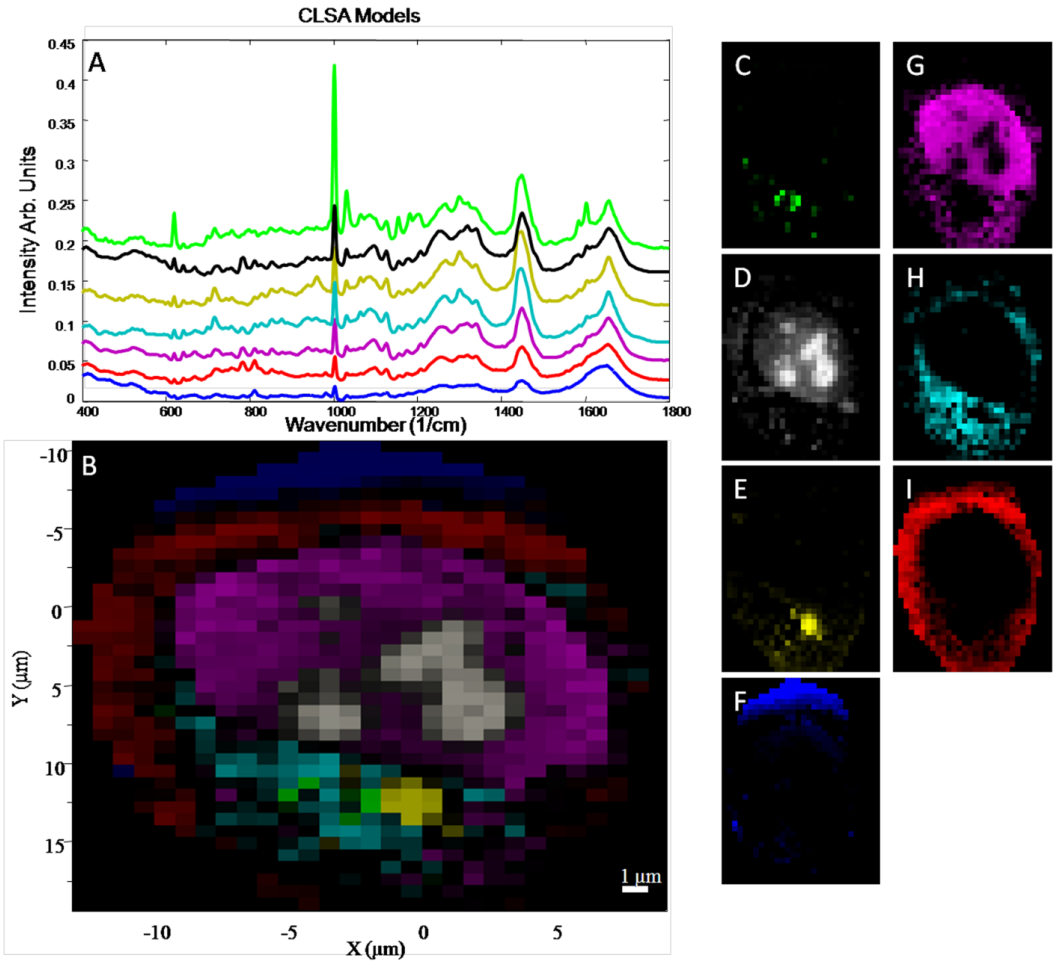


Figure 6.3.: Clustering of spectra identified by unsupervised CLSA. (A) Spectral models generated from the analysis protocol and used to generate the clustered map shown in (B). The right panel (C-I) shows the distribution of each model created in the map. Of particular note, model 1(C), model 6(D) and model 7(H) have strong contributions of the spectra of polystyrene, RNA and lipid respectively. The spectra in (A) are colour coded and correspond to images (B – F), with the exception of Model 6 which corresponds to the white image in (D).

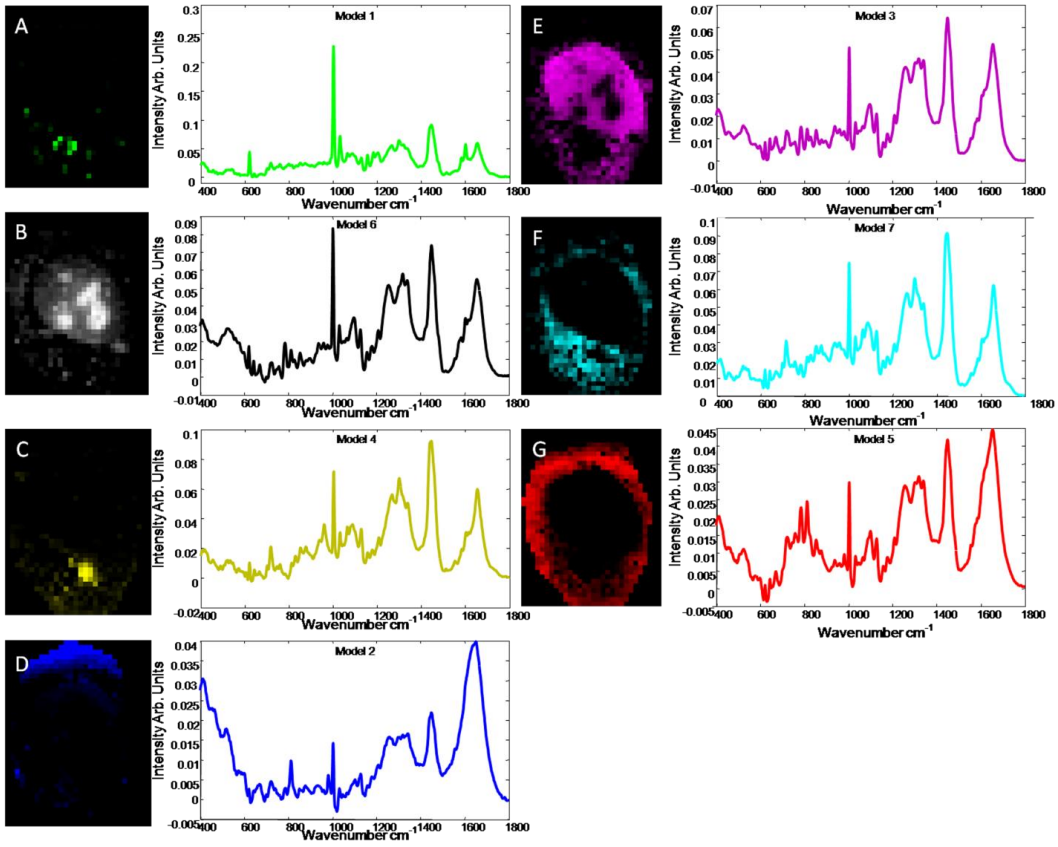


Figure 6.4: A closer look at the generated model spectra created by CLSA (A-G). The overlap between pixels corresponds to a percentage contribution from each particular model. In some instances a pixel may contain 50% of one model and 50% of another, which is highlighted somewhat by the intensity of the pixel, although this is visually subjective.

6.4.3 Simulated Data - Supervised CLSA

Unsupervised CLSA is clearly a powerful technique to analyse the subcellular structure and to identify the presence and distribution of nanoparticles. However, it should be noted that the technique does not yield pure spectra of the components (compare for example Figure 6.4 A with the pure spectrum of polystyrene in Figure 6.1 A), and the respective models are mixtures of spectral signatures of the components and the background cellular spectrum. A secondary approach to CLSA which provides a more supervised approach was therefore also tested. In a

similar way, the simulated datasets were used to assess the technique prior to testing the real Raman cellular map.

The simulated data sets generated to test the unsupervised factor analysis algorithm model generation approach to CLSA were used again to test the supervised approach which uses component spectra of polystyrene, RNA and lipid as the model spectra to generate scores for each spectrum in the data set. In the simulated data shown in Figure 6.5., it is observed that it is possible to identify a trend similar to that seen in Figure 6.2. for the unsupervised CLSA. For RNA and lipid, the trend matches well the predicted response for concentrations as low as 1:0.1, whereupon it deviates from linearity, falling to zero at a ratio of~1:0.03. However, for polystyrene, although the trends are similar, the results are deviate from the predicted response much earlier than the unsupervised CLSA. This indicates that the identification of the components using a supervised CLSA approach may not be as accurate as the model generation approach shown in Figure 6.2. Thus, to test this prediction and for comparison, supervised CLSA was carried out on the same cellular data set using polystyrene, RNA and lipid spectra as the cellular components used to generate the scores for CLSA.

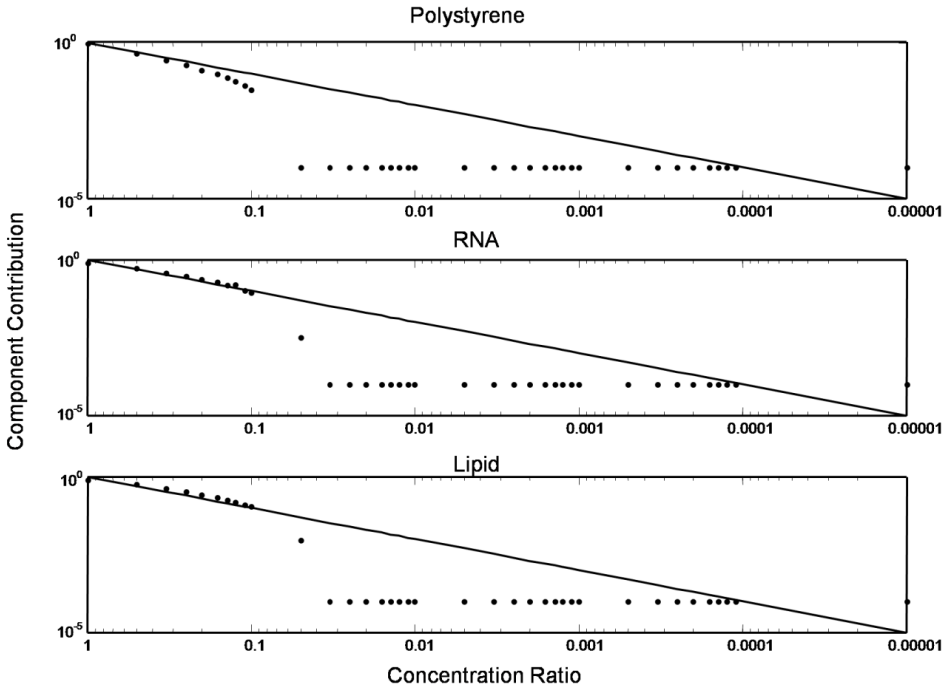


Figure 6.5. Supervised CLSA of simulated spectral data sets of nano-polystyrene, RNA and lipid. In each graph, either the pure spectrum of polystyrene, RNA or lipid was used to calculate the CLSA score. This score was then plotted against the concentration ratio of pure component spectrum: cellular spectrum used to generate the simulated data set.

6.4.4 Single Cell Data - Supervised CLSA

In order to compare the different CLSA approaches, the cellular Raman data set was screened using three pure component spectra individually, nano-polystyrene, RNA and lipid. The aim was to use these spectra to generate the CLSA scores and thus identify regions of the cell which correspond to each spectrum, identifying different regions of the cell based on their biochemical composition and also where the nanoparticles were situated.

The spectra and corresponding score maps are shown in Figure 6.6. A – C. Figure 6.6 A shows a spectrum of polystyrene which was used to screen the map and corresponding visual image of the distribution of nano-polystyrene in the cell. In the image, it is observed that the polystyrene is present in every

spectrum in the cell, albeit in differing amounts based on the pixel intensity at each point. This is not consistent with the model generated CLSA above or with previously published data which show the polystyrene to be localised in clusters surrounding the nucleus[1]. However, the regions of high intensity most likely correspond to the areas which contain the nanoparticles.

Similarly this method for assessing the distribution of RNA and lipids in the cell does not quite reproduce the results observed above for CLSA using the unsupervised factor analysis algorithm. Again, it is observed that the distribution of lipid and RNA is throughout the Raman map of the cell, which, while more plausible for lipids, does not make biological sense for the RNA. Therefore, again it must be concluded that the supervised CLSA approach is prone to error, although it is still possible to compare regions of high intensity to the output of the unsupervised CLSA images above. An arbitrary threshold can be applied to the dataset, as is shown for the three component spectra in the right hand panels of Figure 6.6 A-C. Using this method, the spatial distributions of the components matches well that of the unsupervised CLSA. However this threshold is ambiguous and it is not possible to say from the simulated data at what value an accurate representation of the biochemical distribution in the cell is achieved.

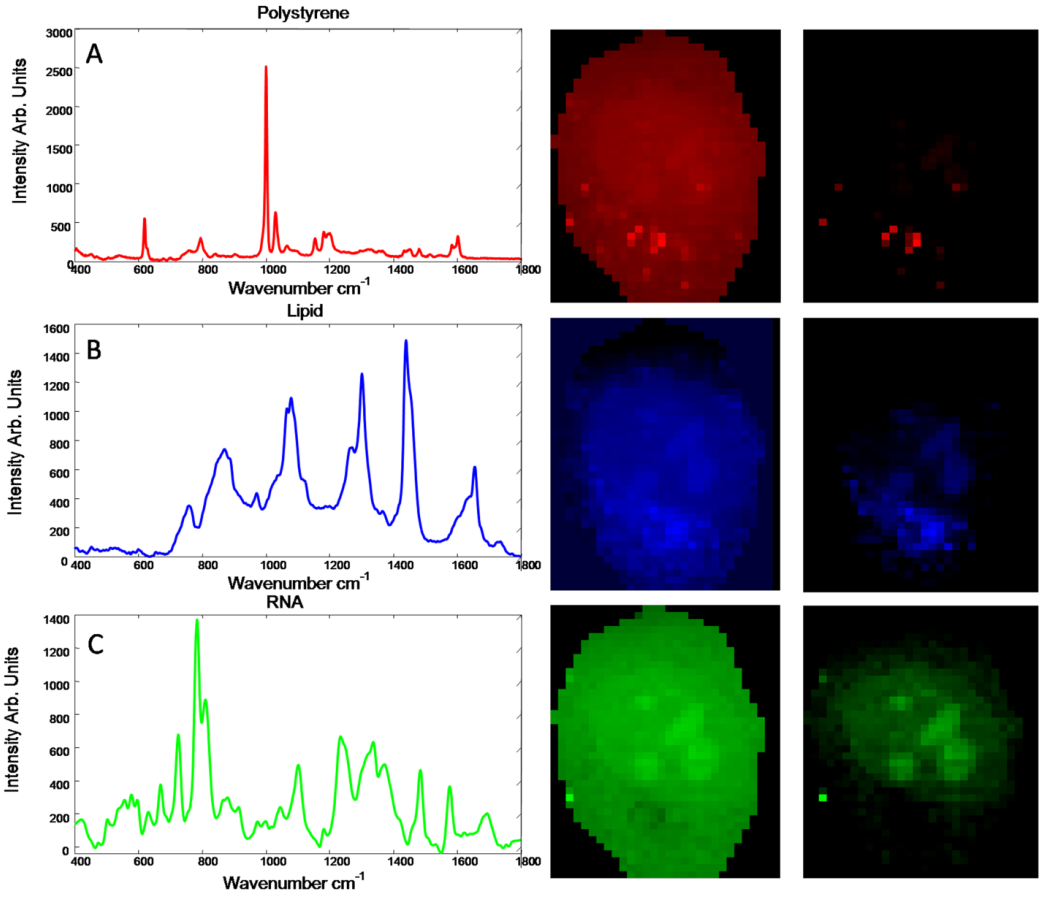


Figure 6.6: Supervised CLSA using component spectra of polystyrene (A), RNA (B) and (C) 3-sn-phosphatidyl ethanolamine. The spectrum of each pure component is shown on the left of the graph, with the corresponding to non-thresholded data shown in the middle and arbitrarily thresholded data shown on the right.

6.4.5 Simulated data –Spectral Cross Correlation Analysis

The observations in Figure 6 that supervised CLSA contained a high level of error in the Raman images prompted a search for an alternate supervised approach to screening Raman data sets which could be used to unambiguously identify regions of the cell which correspond to the pure component spectrum of interest chosen, be that polystyrene, RNA, lipid or any other spectral signature which may be of interest. A novel technique was thus investigated for the analysis of Raman

maps, which uses cross correlation as a method to investigate the presence or absence of a component in a complex Raman data set in a supervised manner. Thus, SCCA was used to screen the same simulated and real data sets for the presence of polystyrene, RNA and lipid for comparison which both methods of CLSA.

Spectral cross correlation analysis (SCCA) was initially investigated using the same simulated data sets that were used to investigate both CLSA approaches. Similar to the supervised CLSA approach, pure component spectra were used to screen each data set for the presence of each in their respective simulated data set. Figure 6.7 compares the results of the simulated SCCA for each of the different components polystyrene, lipid and RNA. In all cases, a correlation of the SCCA co-efficient and the true concentration ratios is observed, but to varying degrees of accuracy.

For polystyrene, a minimum correlation coefficient value of ~ 0.3 is reached at a concentration ratio of cellular: polystyrene spectrum of ~ 1:0.1. This indicates that at this concentration ratio, the presence of the polystyrene spectral fingerprint cannot be distinguished from the cellular spectrum. Thus, for the practical implications of screening a cell for polystyrene nanoparticles, correlation coefficient values at or below 0.3 represent the cellular peaks which overlap with characteristic polystyrene peaks and thus values below this are deemed not to be nanoparticles. This hypothesis was tested using a blank Raman map which contained no polystyrene data in (data not shown) and a value of correlation of 0.3125 was determined, which is close to the predicted value in the simulated data sets. This indicates the need to threshold cellular data in order to identify polystyrene nanoparticles in the cell.

A similar performance was observed for both RNA and lipid simulated data sets, where an initial decrease in the correlation coefficient was observed in relation to concentration ratio of pure component: cell spectrum. Again a minimum baseline correlation coefficient was observed for both RNA and lipid simulated SCCA data. Notably, however, this value was different, in both cases higher, than that observed for polystyrene, possibly due to an increased overlap of Raman bands present in the lipid and RNA spectra with cellular Raman bands in comparison to the polystyrene spectrum. In the case of the lipid contribution, the correlation with the predicted response is quantitatively poor even at ratios above 1:0.1. However, this can possibly be explained by lipid contributions already present in the cellular spectrum and/or the relatively broad lipid bands present in the lipid spectrum used.

The next step was to investigate the performance of SCCA in a real Raman data set of the cell. Thus the previous map was screened in a supervised manner to investigate if nano-polystyrene could be identified in the Raman map. Additionally, the lipid spectrum was used to see if the local cell environment could be investigated. Also, as used in the above supervised CLSA, RNA was used to see if a differentiation could be made between the nucleus and nucleolus.

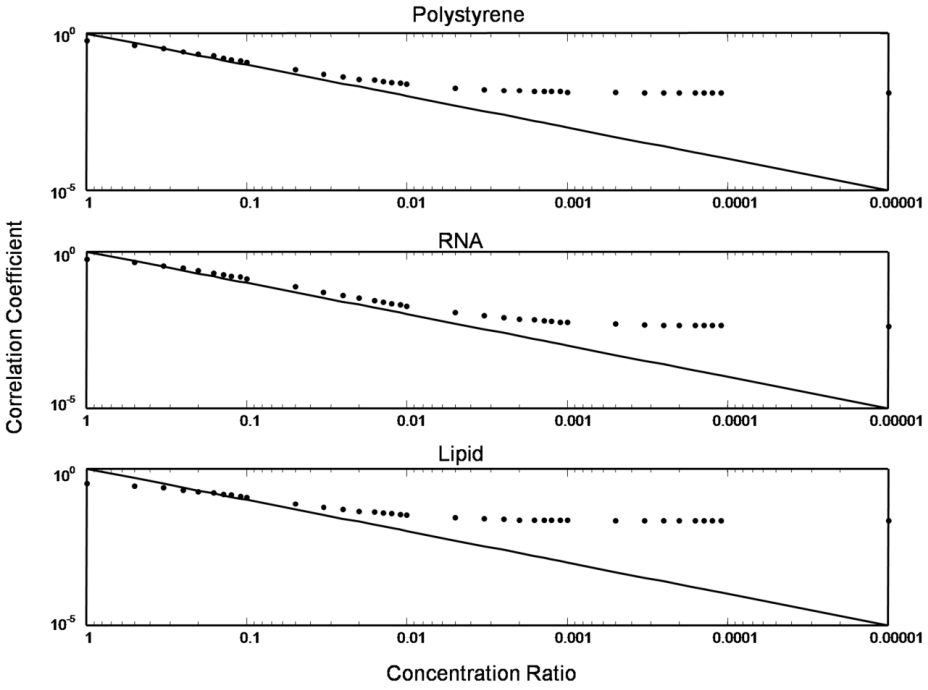


Figure 6.7. SCCA carried out on simulated data sets containing added polystyrene, RNA and lipid component spectra. In each instance, a pure component spectrum of polystyrene, RNA and lipid was cross correlated against each data set to investigate the performance of the technique. The solid line shows the idealised response.

6.4.6 Single Cell Data –SCCA

SCCA was used to screen the Raman data set for the presence of polystyrene, RNA and lipid distributions. The spectra and correlation maps are shown in Figure 6.8. In figure 6.8 A, the spectrum of polystyrene is shown in red and the corresponding correlation map is shown adjacent for both thresholded (right) and non-thresholded (left) datasets. This map shows the distribution of polystyrene nanoparticles in the Raman map. Importantly, the threshold which was predicted from the simulated data, or more simply from a cross-correlation of the component spectrum with the raw average cellular spectrum, was applied to the data set and returned a map which corresponded to the previously observed Raman image from the unsupervised CLSA (Fig 6.3 A). Notably, however, the

spectrum is the pure spectrum of polystyrene, rather than a cellular/polystyrene mixture. This result shows the capability for a supervised approach for the unambiguous identification of polystyrene nanoparticles in complex Raman spectroscopic data sets.

Furthermore, to investigate how SCCA can be used to probe the local cellular environment, the lipid spectrum was used to screen the data set (Fig 6.8 B). Again applying a threshold to the data set it is possible to identify regions of the cell which contain a high density of lipids using a supervised approach to Raman analysis. Thus it is possible to investigate the local cell environment to which the nanoparticles are trafficked after 24hrs. This is consistent with the previous K-means cluster analysis[1] which suggests that indeed the nanoparticles are located in a highly lipid rich environment.

As an additional demonstration of the potential of SCCA, a pure RNA spectrum was cross correlated against the data set to see if it was possible to differentiate spectra which corresponded to the nucleolus of the cell and thus differentiate between DNA and RNA rich regions of the cell. Figure 4.5.7.1. C shows that it is possible to identify the nucleolus of the cell using cross correlation analysis. It was also observed that a high correlation coefficient was present in regions outside the nucleus. This could possibly correspond to cytoplasmic ribosomal RNA (rRNA) or cytoplasmic messenger RNA (mRNA). Thus a novel approach for extracting complex spectral information from Raman data sets is demonstrated in SCCA.

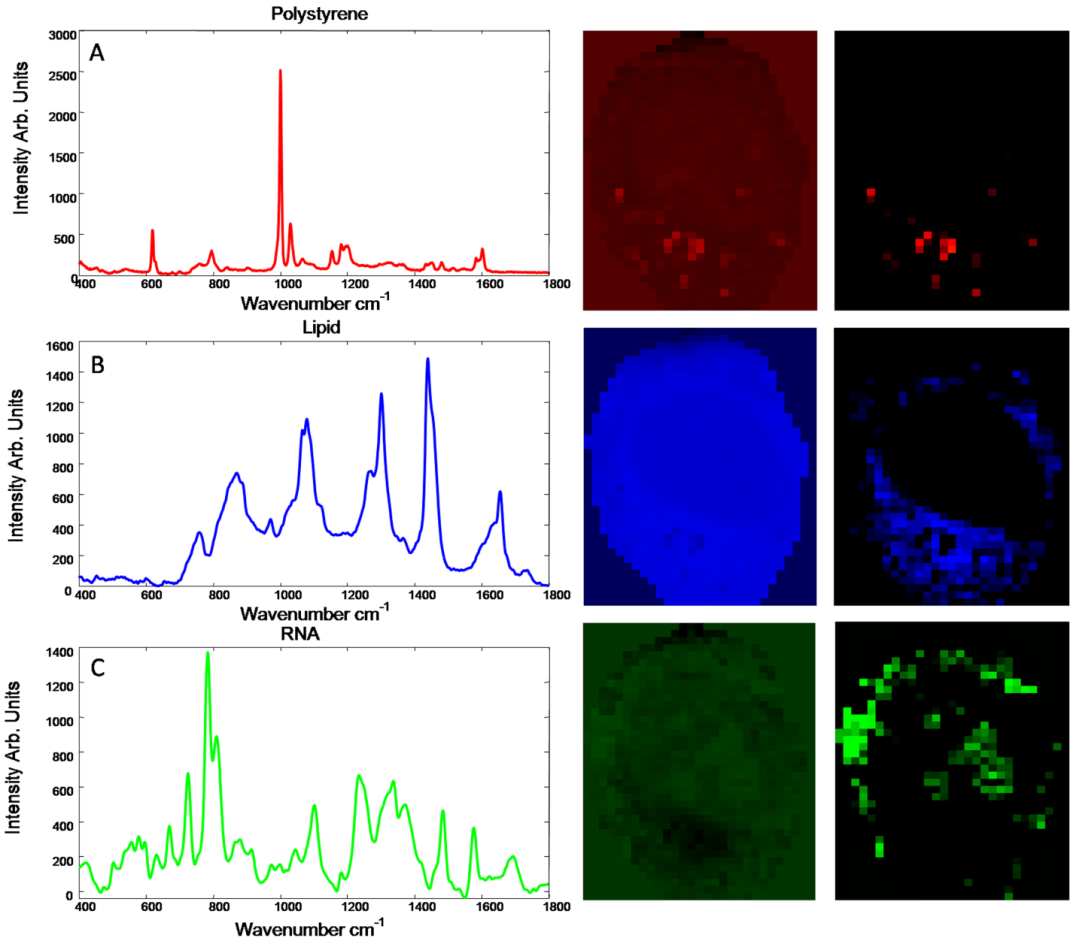


Figure 6.8: SCCA analysis using component spectra of polystyrene (A), 3-sn-phosphatidyl ethanolamine (B) and RNA (C). The spectrum of each pure component is shown on the left of the figure and the correlation maps for non-thresholded shown in the middle and thresholded on the right.

6.5 Discussion

Raman spectroscopy is a powerful tool for the investigation of biological samples. Previous studies have shown the capability of the technique to investigate sub cellular structures and processes which provide Raman images comparable to images observed using wide-field and confocal fluorescent microscopy[1,36,45–47]. Notably, however, Raman spectroscopy is a label free method which provides a visualization of the biochemical make up of a cell without costly and time consuming processing with reagents, and when combined with appropriate

analysis methods can provide a wealth of information pertaining to biological processes in the cell. The aim of this paper was thus to investigate two analytical approaches both in an unsupervised and supervised approach and assess their ability to identify polystyrene nanoparticles and biochemical distributions in a single cell Raman map.

Unsupervised CLS analysis is demonstrated to be capable of identifying the presence of nanoparticles in regions of the cell. However, while this method is valuable for identifying distributions in the cell, the model spectra generated in this manner must be further analysed to extract any real biochemical information. Therefore, while the analysis of the simulated dataset in figure 6.2. indicates that the unsupervised model has a higher accuracy, the model spectra yielded by the unsupervised CLS analysis do not directly compare to the pure component spectra shown in Figure 6.1 and therefore cannot be used to unambiguously identify the contributing components.

In contrast, employing supervised approaches to the analysis of Raman data sets allows for the spectral array to be screened directly with the nanoparticle or pure biochemical component spectrum of interest. Analysis in this way enables a direct screening of the cellular distribution of a particular component while simultaneously probing the chemical or biochemical environment of the particular location in the cell. CLSA and SCCA are both used in a supervised approach for analysing Raman cellular data sets (Figure 6.6 and Figure 6.6). However, unthresholded, both show a degree of error for all three components tested (nano-polystyrene, RNA and Lipids). To correct for this, a threshold can be applied to both CLSA and SCCA. Importantly, this threshold should not be applied in an arbitrary manner, as this facilitates a loss of information from the

dataset. While thresholding for supervised CLSA is arbitrary and subjective, the simulated datasets generated for SCCA provided a good estimation of where this thresholding should take place and in combination with cellular data containing no nanoparticles it was possible to accurately reveal where the nanoparticles were located in the cell. It should be noted that the thresholding level appears to be dependent on the spectral profile of the individual component, as it is dependent on the degree of similarity of the spectrum of the target component with that of the environment. Incorrect correction of spectral background may also add to the threshold. On the other hand the simulated data for supervised CLSA did not provide a threshold value to apply to the dataset and thus was arbitrarily thresholded, which is far from ideal to gain any reliable information about the dataset. Therefore, SCCA provides a more reliable supervised approach for identification of nanoparticles and other biological components when used in combination with a threshold generated by simulated datasets. In addition, quantitative information can be extracted from the simulated data sets, with each of the three approaches showing some level of quantification based on how well the matched the predicted response, with SCCA showing the highest level of sensitivity of the three techniques. SCCA is specifically a supervised approach, as it is necessary to provide the pure component spectrum. However, it is conceivable the technique could be extended to a library of reference spectra which could in turn be screened against the data set in an unsupervised manner.

6.6 Conclusions

CLSA and SCCA are shown to be two methods capable of identifying intracellular polystyrene nanoparticles and also to probe the local biochemical

environment the nanoparticles are trafficked to within the cell. CLSA is a relatively straight forward method for analysing spectroscopy data sets. However, SCCA is demonstrated in the simulated data sets to be a more sensitive approach for nanoparticle identification. It is envisaged that both these and other supervised methods will provide analytical approaches which can be used not only as identification methods for other nanoparticles inside cells and detection of resultant biochemical changes, but also to provide alternate analytical approaches to the study of other processes such as chemotherapeutic response of cells to drugs. Additionally the full quantitative nature of these analytical approaches will need to be explored if Raman spectroscopy is to become a routine application in the study of nano-bio interactions and beyond.

Acknowledgements: This research was supported by the Integrated NanoScience Platform, Ireland (INSPIRE), and the National Biophotonics and Imaging Platform (NBIP) Ireland, both funded under the Higher Education Authority PRTLI (Programme for Research in Third Level Institutions) Cycles 4 and 5, co-funded by the Irish Government and the European Union Structural fund.

6.7 References

1 J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers and H. J. Byrne, *Analyst*, 2012, **137**, 1111–9.

2 S. Dhar and N. Kolishetti, *Proc. Natl. Acadamy Sci.*, 2011, **108**, 1850–1855.

3 Y. Wang, Y. Wang, J. Xiang and K. Yao, *Biomacromolecules*, 2010, **11**, 3531–8.

4 T. Lammers, F. Kiessling, W. E. Hennink and G. Storm, *J. Control. Release*, 2012, **161**, 175–87.

5 A. Kumar, P. K. Vemula, P. M. Ajayan and G. John, *Nat. Mater.*, 2008, **7**, 236–41.

6 I. Perelshtein, G. Applerot, N. Perkas, J. Grinblat and A. Gedanken, *Chemistry*, 2012, **18**, 4575–82.

7 H. Yan, H. S. Choe, S. Nam, Y. Hu, S. Das, J. F. Klemic, J. C. Ellenbogen and C. M. Lieber, *Nature*, 2011, **470**, 240–4.

8 S. S. Kelkar and T. M. Reineke, *Bioconjug. Chem.*, 2011, **22**, 1879–903.

9 S. Wang, G. Kim, Y.-E. K. Lee, H. J. Hah, M. Ethirajan, R. K. Pandey and R. Kopelman, *ACS Nano*, 2012.

10 J.-P. Colinge, C.-W. Lee, A. Afzalian, N. D. Akhavan, R. Yan, I. Ferain, P. Razavi, B. O'Neill, A. Blake, M. White, A.-M. Kelleher, B. McCarthy and R. Murphy, *Nat. Nanotechnol.*, 2010, **5**, 225–9.

11 T. R. Downs, M. E. Crosby, T. Hu, S. Kumar, A. Sullivan, K. Sarlo, B. Reeder, M. Lynch, M. Wagner, T. Mills and S. Pfuhler, *Mutat. Res.*, 2012, **745**, 38–50.

12 M.-F. Song, Y.-S. Li, H. Kasai and K. Kawai, *J. Clin. Biochem. Nutr.*, 2012, **50**, 211–6.

13 B. Ziemba, A. Janaszewska, K. Ciepluch, M. Krotewicz, W. a Fogel, D. Appelhans, B. Voit, M. Bryszewska and B. Klajnert, *J. Biomed. Mater. Res. A*, 2011, **99**, 261–8.

14 R. P. Singh and P. Ramarao, *Toxicol. Lett.*, 2012, 1–11.

15 X. Zheng, J. Tian, L. Weng, L. Wu, Q. Jin, J. Zhao and L. Wang, *Nanotechnology*, 2012, **23**, 055102.

16 A. Jaeger, D. G. Weiss, L. Jonas and R. Kriehuber, *Toxicology*, 2012, **296**, 27–36.

17 G. Hunt and M. Riediker, *Nanotechnol. Perceptions*, 2011, **7**, 82–98.

18 H. J. Byrne, I. Lynch, W. H. De Jong, W. G. Kreyling, S. Loft, M. V. D. Z. Park, M. Riediker and D. Warheit, 2008, 1–30.

19 P. Sandin, L. W. Fitzpatrick, J. C. Simpson and K. A. Dawson, *ACS Nano*, 2012, **6**, 1513–21.

20 F. Fazlollahi, S. Angelow, N. R. Yacobi, R. Marchelletta, A. S. L. Yu, S. F. Hamm-Alvarez, Z. Borok, K.-J. Kim and E. D. Crandall, *Nanomedicine*, 2011, **7**, 588–94.

21 E. Jan, S. J. Byrne, M. Cuddihy, A. M. Davies, Y. Volkov, Y. K. Gun'ko and N. A. Kotov, *ACS Nano*, 2008, **2**, 928–38.

22 T. Y. Ohulchanskyy, I. Roy, K.-T. Yong, H. E. Pudavar and P. N. Prasad, *Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol.*, 2010, **2**, 162–75.

23 G. D. Byrne, M. C. Pitter, J. Zhang, F. H. Falcone, S. Stolnik and M. G. Somekh, *J. Microsc.*, 2008, **231**, 168–79.

24 J. Contreras, J. Xie, Y. J. Chen, H. Pei, G. Zhang, C. L. Fraser and S. F. Hamm-Alvarez, *ACS Nano*, 2010, **4**, 2735–2747.

25 T. Tenuta, M. P. Monopoli, J. Kim, A. Salvati, K. a Dawson, P. Sandin and I. Lynch, *PLoS One*, 2011, **6**, e25556.

26 G. Wang, A. S. Stender, W. Sun and N. Fang, *Analyst*, 2010, **135**, 215–21.

27 J. Zhou, C. Leuschner, C. Kumar, J. Hormes and W. O. Soboyejo, *Mater. Sci. Eng. C*, 2006, **26**, 1451–1455.

28 A. M. Schrand, J. J. Schlager, L. Dai and S. M. Hussain, *Nat. Protoc.*, 2010, **5**, 744–57.

29 K. Shapero, F. Fenaroli, I. Lynch, D. C. Cottell, A. Salvati and K. a Dawson, *Mol. Biosyst.*, 2011, **7**, 371–8.

30 M. Davoren, E. Herzog, A. Casey, B. Cottineau, G. Chambers, H. J. Byrne and F. M. Lyng, *Toxicol. In-Vitro*, 2007, **21**, 438–48.

31 F. M. Lyng, E. O. Faoláin, J. Conroy, a D. Meade, P. Knief, B. Duffy, M. B. Hunter, J. M. Byrne, P. Kelehan and H. J. Byrne, *Exp. Mol. Pathol.*, 2007, **82**, 121–9.

32 K. M. Ostrowska, A. Malkin, A. Meade, J. O'Leary, C. Martin, C. Spillane, H. J. Byrne and F. M. Lyng, *Analyst*, 2010, **135**, 3087–93.

33 H. Nawaz, F. Bonnier, P. Knief, O. Howe, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2010, **135**, 3070–6.

34 H. Nawaz, F. Bonnier, A. D. Meade, F. M. Lyng and H. J. Byrne, *Analyst*, 2011, **136**, 2450–63.

35 F. Bonnier, P. Knief, B. Lim, A. D. Meade, J. Dorney, K. Bhattacharya, F. M. Lyng and H. J. Byrne, *Analyst*, 2010, **135**, 3169–77.

36 F. Bonnier, a D. Meade, S. Merzha, P. Knief, K. Bhattacharya, F. M. Lyng and H. J. Byrne, *Analyst*, 2010, **135**, 1697–703.

37 P. Knief, C. Clarke, E. Herzog, M. Davoren, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2009, **134**, 1182–91.

38 K. Kneipp, A. S. Haka, H. Kneipp, K. Braizadegan, N. Yoshizawa, C. Boone, K. E. Shafer-Peltier, J. T. Motz, R.R. Dasari and M.S. Feld. *Appl. Spectrosc.*, 2002, **56**, 150–154.

39 J. Kneipp, H. Kneipp, M. McLaughlin, D. Brown and K. Kneipp, *Nano Lett.*, 2006, **6**, 2225–31.

40 M. Hedegaard, C. Matthäus, S. Hassing, C. Krafft, M. Diem and J. Popp, *Theor. Chem. Acc.*, 2011, **130**, 1249–1260.

41 M. MiljkoviÄ‡, T. Chernenko and M. Romeo, *Analyst*, 2010, 2002–2013.

42 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

43 F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 322–32.

44 C. L. Zavaleta, B. R. Smith, I. Walton, W. Doering, G. Davis, B. Shojaei, M. J. Natan and S. S. Gambhir, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 13511–6.

45 C. Matthäus, T. Chernenko, J. a Newmark, C. M. Warner and M. Diem, *Biophys. J.*, 2007, **93**, 668–73.

46 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

47 K. Klein, A. M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F. Jamitzky, G. Morfill, R. W. Stark and J. Schlegel, *Biophys. J.*, 2012, **102**, 360–8.

Chapter 7: Conclusions

Given the drive for a reduction in the use of animal models for evaluating toxicity, screening of drugs and even cosmetics, due to regulatory developments in both the EU and US (EU Directive-2010/63/EU and US Public Law 106-545, 2010, 106th Congress)[1–3] generally based on the 3 R's of Russell and Burch[30] to replace, reduce and refine the use of animals used for scientific purposes, there is increased emphasis on the development of reliable and rapid *in-vitro* screening methodologies. This includes more representative culture models which better mimic the *in-vivo* environment as well as more rapid, cost efficient, high content, and ideally label free screening technologies. It is crucial, however, that these models and technologies are well validated against established gold standards[4,5].

Raman spectra, in principle, contain high content information about the biochemical make-up of the sample, and changes to it, related to pathology or an external agent. Raman spectra contain numerous peaks which vary dependently and independently of each other. Chapter 2 reviewed some of the current and emerging applications of Raman spectroscopy in the field of Nanomedicine, for example. Extraction and analysis of the relevant data requires the application of multivariate statistical protocols, and validation of such protocols used in vibrational spectroscopy, many of the commonly employed modes of which were introduced in Chapter 3, is essential to ensure that the spectral analysis is reliable and accurate for standardised and routine usage. In the development of novel *in-vitro* screening tools for both nano and pharmacological screening, it is therefore imperative that sample preparation protocols, instrumental reliability and

multivariate routines are robust and reproducible, to ensure a smooth transition from laboratory bench to clinical and industrial settings.

Focusing on the multivariate statistical analysis, previously published work by Bonnier and Byrne, 2012[6], aimed to elucidate the use and interpretability of PCA in vibrational spectral applications, using both real and simulated data. Similarly, validation and development of the RMie-EMSC algorithm was done using simulated and real data in the work of Bassan et al 2010[7]. These selected examples show the applicability of simulations in validation of multivariate analysis as well as spectral prepossessing in a biomedical vibrational spectroscopic context.

Crucial to the application of Raman spectroscopy in these areas is the use of data mining and data analysis, as a way to identify trends in the spectral data, which is important as a tool to extract and distil spectral information which may not be apparent to the eye.

For diagnostics, classification of samples is sufficient, and accuracy of sensitivities and specificities are important for translational to the clinic. To go beyond, and make use of analytical capabilities of Raman spectroscopy, data mining is important in the exploration of the potential of the technique for dose dependent in-vitro studies and mechanistic responses for both drug and nano-toxicological and screening applications. It is also imperative that spectroscopic techniques are compared and validated against gold standard assays and diagnostic classification protocols to ensure accurate, reliable and robust spectral methodologies in these settings.

A differentiation should be made between the screening of cell populations i.e. point by point cellular acquisition and Raman as a whole cell and sub-cellular imaging, technique. Both provide label free analysis, although the type of knowledge gained is different; cell population screening provides information how a drug or nanomaterial may effect the average cell viability of a population [8–10], while as an imaging technique, information is gained about the subcellular spatial distribution and mode of action of drugs and nanomaterials[11,12].

Crucially, for real applications and particularly in the instance of drug interactions, it is difficult to tell whether these differences are inherently based on cell to cell variability or whether they are dependent on the primary action of the drug (i.e. the direct chemical effects) or the secondary effects the drug has on the cell (i.e. the response of the cell to said drug).

In Chapter 4, simulated datasets were used to evaluate the capability of PLSR to extract known and systematic spectral variation from a control dataset, which contained intrinsic experimental variability. The spectral variations introduced varied linearly with the applied drug dose and also with the cell population response, as measured by a standard cytotoxicity assay. Notably, however, the two spectral variations are not completely independent, as the viability response is sigmoidal dependent on the applied dose.

In the case where only a concentration dependent systematic variation in the spectra is introduced, the PLSR model provides an accurate predictive response tool, the regression co-efficients of which are based on the systematic variation which has been introduced to the dataset, linearly dependent on the targets. The model shows high sensitivity, and the limits of detection are determined only by the intrinsic variability of the experimental method, as

determined by the PLSR of the Control spectral dataset. This limit can be improved by optimising sample preparation and measurement protocols. In principle, such a PLSR model can predict the response of a drug dose in a cell population, or determine an unknown drug dose from a measured spectral response.

However, the spectral changes which result from the interaction and action of a drug within a cell are manifold, and it is of interest to differentiate the spectral signatures of the direct interaction from the subsequent cellular response. Notably, this study demonstrates that, although PLSR predictive models based on regression of the combined dataset, including all spectral responses, against the target of concentration range produce a similarly accurate, linear predictive model, the contributing regression co-efficient (RCs) are derived exclusively from the introduced concentration dependent variations in ranges where all other spectral variations are limited. For example, as shown in Figures 4.8 and 4.9, regression over the limited range of C+4 produces a model which is based on RCs which include contributions derived from the direct effect of the interaction of the drug within the cell (Concentration construct), as well as the resultant cytological response (Viability construct). Thus, care should be taken in interpreting the spectral features which contribute to such regressions to elucidate the underlying mechanisms.

Nevertheless, in the sub-lethal regions, the direct effects of the drug interaction can confidently be investigated employing such a PLSR analysis of Raman spectral data, independent of the cytological responses, and these are easily discernible above the intrinsic variability of the control. Although this seems a trivial conclusion, such rapid, label free analysis could prove invaluable

in screening of, for example, the mechanisms and efficacy of drug interactions, evaluating drug uptake and receptor binding or nanoparticle uptake and trafficking in regions where cytoxicity assays are insensitive.

The use of a parallel cytotoxic assay, such as MTT, serves as a range finding test to establish the IC_{50}, but also provides vital information about the sub-lethal doses and maximum responses. It also provides a target for regression of the data in the regions of toxicity. Thus, the subsequent cytological effects can be differentiated from the direct chemical effects of the agent and extracted from the overall spectral response in the dose range where the viability is impacted, and the cellular response can be independently mapped spectroscopically, as a function of dose and time. Notably, the model described in Chapter 4, which includes a single spectral construct to represent the cellular response is very simplistic, as the response is a cascade of many responses, depending on the mechanism of interaction[13]. Nevertheless, the analysis presented here demonstrates that the spectral fingerprints of the direct mechanisms of interaction and the subsequent cellular responses can be independently extracted from the dose dependent spectral data, and thus, ultimately with improved screening sensitivities and speeds, Raman spectroscopy could be employed to monitor in quasi real time, in a label free manner, the efficacy and mode of action of, for example chemotherapeutic agents and other exogenous agents, laying the basis for improved quantitative structure activity relationships to guide drug development or chemical regulation strategies.

This study demonstrates the reliability and also limitations of PLSR as a method for predictive modelling and analysis of spectroscopic signatures of cellular responses to exogenous agents such as radiation, chemotherapeutic

agents or toxins. The spectroscopic profiles at any dose/time point can derive from a complex mixture of direct interactions within the cell and a cascade of subsequent cellular response. The analysis demonstrates that care should be taken in choosing the response range and also highlights the importance of parallel cytological assays in guiding the modelling and analysis. Correct choice of range can help differentiate between the signatures of direct interactions, which are dominant at sub-lethal doses and those of the subsequent cellular response which evolve with increasing dose.

PCA of the same simulated spectral dataset was also investigated for its ability to extract the systematic and continuously variable spectral perturbations introduced. Limitations of PCA of the data are shown in figure 5.3 A, whereby the algorithm was unable to extract the desired spectral features from the dataset, as the magnitude of the perturbation was less than the intrinsic variability of the cellular data. A successful partition of the data is shown to be possible when the algorithm is seeded with the known spectral variation, as demonstrated in figure 5.5 A and B. In Dataset 3, which is continuously perturbed by the addition of weighted contributions of the two spectral constructs, seeding the dataset with the minority perturbation enables the continuous differentiation of the data, and extraction of both independent spectral perturbations. Further improvements in separation are shown using 1^{st} and 2^{nd} derivative spectral data for the seeded datasets such that, in the case of the SePCA of Dataset 3, the PCA scatter plot shown in figure 5.20A reproduces to some degree the experimental dose dependent toxicity study of Nawaz et al. This has implications for *in-vitro* spectral screening platforms as it shows that the correct trend in simulation can be extracted from the data once the correct features are described to the algorithm

i.e. a seeded approach. This is a positive step towards a multivariate dose response curve.

The improvements shown have possible ramifications for both diagnostics and *in-vitro* screening. Notably, however, in comparison to the PLSR approach of Chapter 5, the method is supervised, in the sense that it requires some prior knowledge of the spectral changes in the data set. In terms of Construct 1, this could be facilitated by a library of spectral signatures of, for example, DNA major and minor groove binders and intercalators, allowing a rapid screening of mechanisms of action of novel chemotherapeutic agents. In a similar fashion, spectral signatures could be established to represent Adverse Outcome Pathways (AOPs), an approach to representation of toxicology recently endorsed by the OECD[14]. In this approach, while the chemical binding of the agent to the receptors represents the Molecular Initiator Event (MIE), cascade of events leading to, for example apoptosis or necrosis constitute the AOP, which could be represented by distinct spectral signatures.

For diagnostic applications such as classification e.g. using support vector machines (SVM) or linear discriminant analysis (LDA), in which PCA coefficients are input to the algorithms, seeding in combination with 1st and 2nd derivative spectra may provide improvements in dividing the data for training and thus, improvements in the diagnostic classification accuracy if the correct variable features can be identified across the patient data.

The nature of the continuously varying spectral changes is also relevant for the interpretation of experimental changes. In this instance (Dataset 3), the changes are continuous and linearly increasing across the entire dataset. However, in experimental data, the changes may not be present in a continuous or linear

fashion, or across the entire sampled range. If, as in many instances, the loadings contain an ensemble of spectral features, multiple trends may be responsible for the pattern of separation in the data. By seeding with the correct peaks the pattern of partitioning in the data can be more accurately identified and adjusted based on the correct spectral changes in the data.

This study demonstrates an analytical methodology, seeded PCA, which increases the potential of the PCA algorithm to separate spectrally distinct data, particularly in the case where continuous but minor variations are present over a dataset range. The use of 1st and 2nd derivatisation of the dataset is demonstrated to further enhance the differentiation potential of the algorithm. This has important ramifications for improving separation of spectra, with a particular emphasis on biomedical spectroscopy, be that in spectral diagnostics i.e. classification protocols, and/or *in-vitro* screening of drugs and nano-materials. The study also demonstrates the benefits of analysis of simulated datasets in the development and validation of novel multivariate analysis algorithms.

Moving from dose dependent responses in cell populations to analysis of cells and cellular processes, Raman spectroscopy may also be deployed as an imaging technique with subcellular resolution. Previous studies have shown the capability of the technique to investigate sub cellular structures and processes. As an optical technique, Raman images are comparable to those produced using wide-field and confocal fluorescent microscopy[11,15–18]. This is demonstrated further in chapter 6, in which two analytical approaches, unsupervised and supervised where assessed based on their ability to identify polystyrene nanoparticles and biochemical distributions in an experimental single cell Raman

map. Simulation models were employed to quantitatively compare the relative sensitivities of the data mining techniques.

Unsupervised CLS analysis is demonstrated to be capable of identifying the presence of nanoparticles in regions of the cell. However, while this method is valuable for identifying distributions in the cell, the spectra generated in this manner must be further analysed to extract any real biochemical information. Therefore, while the analysis of the simulated dataset in figure 6.2. indicates that the unsupervised model has a higher accuracy, the model spectra yielded by the unsupervised CLS analysis do not directly compare to the pure component spectra shown in Figure 6.1 and therefore cannot be used to unambiguously identify the contributing components.

In contrast, employing supervised approaches to the analysis of Raman data sets allows for the spectral array to be screened directly with the nanoparticle or pure biochemical component spectrum of interest. Analysis in this way enables a direct screening of the cellular distribution of a particular component while simultaneously probing the chemical or biochemical environment of the particular location in the cell. CLSA and SCCA are both used in a supervised approach for analysing Raman cellular data sets (Figure 6.6 and Figure 6.6). However, un-thresholded, both show a degree of error for all three components tested (nano-polystyrene, RNA and Lipids). To correct for this, a threshold can be applied to both CLSA and SCCA. Importantly, this threshold should not be applied in an arbitrary manner, as this facilitates a loss of information from the dataset. While thresholding for supervised CLSA is arbitrary and subjective, the simulated datasets generated for SCCA provided a good estimation of where this thresholding should take place and, in combination with cellular data containing

no nanoparticles, it was possible to accurately reveal where the nanoparticles were located in the cell. It should be noted that the thresholding level appears to be dependent on the spectral profile of the individual component, as it is dependent on the degree of similarity of the spectrum of the target component with that of the environment. Inappropriate correction of spectral background may also add to the threshold. On the other hand, the simulated data for supervised CLSA did not provide a threshold value to apply to the dataset and thus was arbitrarily thresholded, which is far from ideal to gain any reliable information about the dataset. Therefore, SCCA provides a more reliable supervised approach for identification of nanoparticles and other biological components when used in combination with a threshold generated by simulated datasets. In addition, quantitative information can be extracted from the simulated data sets, each of the three approaches showing some level of quantification based on how well they matched the predicted response, SCCA showing the highest level of sensitivity of the three techniques. SCCA is specifically a supervised approach, as it is necessary to provide the pure component spectrum. However, it is conceivable the technique could be extended to a library of reference spectra which could in turn be screened against the data set in an unsupervised manner.

CLSA and SCCA are shown to be two methods capable of identifying intracellular polystyrene nanoparticles and also to probe the local biochemical environment the nanoparticles are trafficked to within the cell. CLSA is a relatively straight forward method for analysing spectroscopy data sets. However, SCCA is demonstrated in the simulated data sets to be a more sensitive approach for nanoparticle identification. It is envisaged that both these and other supervised methods will provide analytical approaches which can be used not only as

identification methods for other nanoparticles inside cells and detection of resultant biochemical changes, but also to provide alternate analytical approaches to the study of other processes such as chemotherapeutic response of cells to drugs.

Previous work investigating biochemical information, which can been gained from using multivariate statistical methods has been explored by Bonnier et al[19] where PCA and k-means were used to investigate cellular data, while this study primarily uses real data to explore the capabilities of these methods, some of the work in this thesis uses simulations to probe the usefulness of such methods in biomedical spectroscopy.

While these methods provide approaches for the study of individual cells, it is important to note that the exact changes present in the spectra are still partially unknown and thus it is difficult to ensure that all spectra are classified and group precisely, in an error free manner.

7.1 Future work

Future work may involve the design of an advanced cellular simulated model, as indicated in figure 7.1. As a demonstration of principle, the generation of a model dataset was initially undertaken using K-Means clustering of the nanoparticle exposed cellular dataset of Dorney et al[11]. The purpose here was to generate 6 clusters in the dataset which corresponded roughly to the different spatial regions of the cell i.e. the nucleus, nucleolus, perinuclear regions of the cell, the surrounding cytoplasm and an external agent, in this case polystyrene nanoparticles. This was achieved by splitting the resulting matrix obtained from KMCA into separate regions which correspond to the clusters identified in the

analysis. The six matrixes were then converted to binary format by replacing all values to a series of ones and zeros. This process allows for the generation of template datasets or images. Next, spectra were chosen which, in this initial dataset, are pure component spectra of two lipid spectra, phosphatidyl inositol and phosphatidyl ethanolamine, DNA, RNA, a spectrum of the background and a spectrum of polystyrene nanoparticles.

These spectra are then used to populate the image templates generated using K-means clustering. Thus, if the templates consist of a series of ones and zeros, by a process of multiplication only, the regions which contain a one will contain a spectrum. Thus it is possible to populate the dataset with any spectrum of interest. As stated, in this case the dataset is based on nanoparticle cell interaction, so in this initial simplified dataset the examples used consist of pure spectral components which are matched to a corresponding regional distribution. This is shown in table 7.1

This outlines a preliminary example of a biochemical spatial simulation of the cell based on pure cellular components. While this example is simplistic in nature, mixing of the base components in different weightings may lead to a better understanding of statistical methods used for Raman cellular imaging. Extending this concept of Raman imaging to high content cellular imaging, such a cellular simulated model could serve as a template to validate and extend data mining, towards a robust analysis protocol.

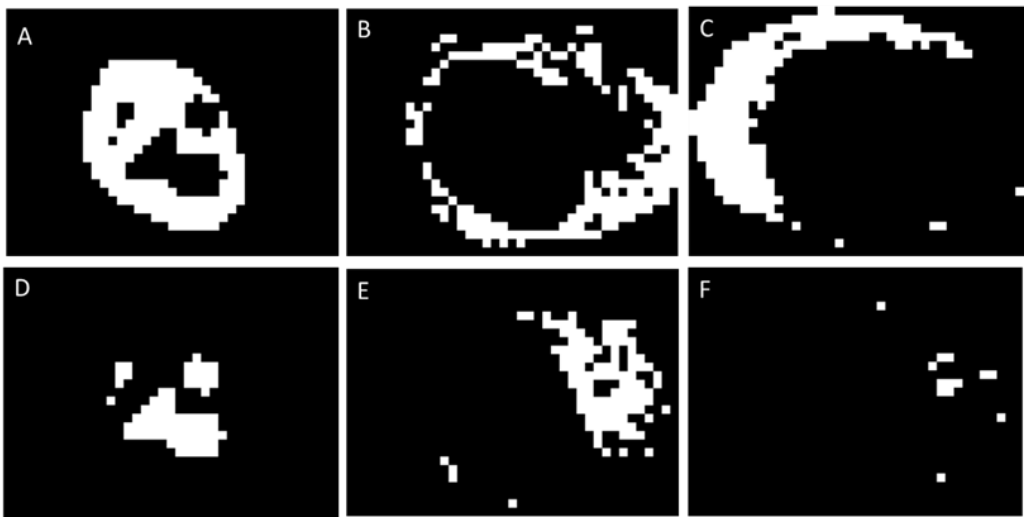


Figure 7.1. Initial template regions showing the known spatial distribution of pure component spectra representing the Nucleus (A), Perinuclear 1 (B), Cytoplasm (C), Nucleolus (D), Perinuclear 2 (E) and Polystyrene nanoparticles (F). This spectral regions correspond to the pure component spectra in table 7.1 and figure 7.2.

Spectrum	Region
Phosphatidyl Inositol	Perinuclear Region 1
Phosphatidyl Ethanolamine	Perinuclear Region 2
RNA	Nucleolus
DNA	Nucleus
Background	Cytoplasm
Polystyrene Nanoparticles	External Agent

Table 7.1. Pure spectral components and corresponding regional distributions

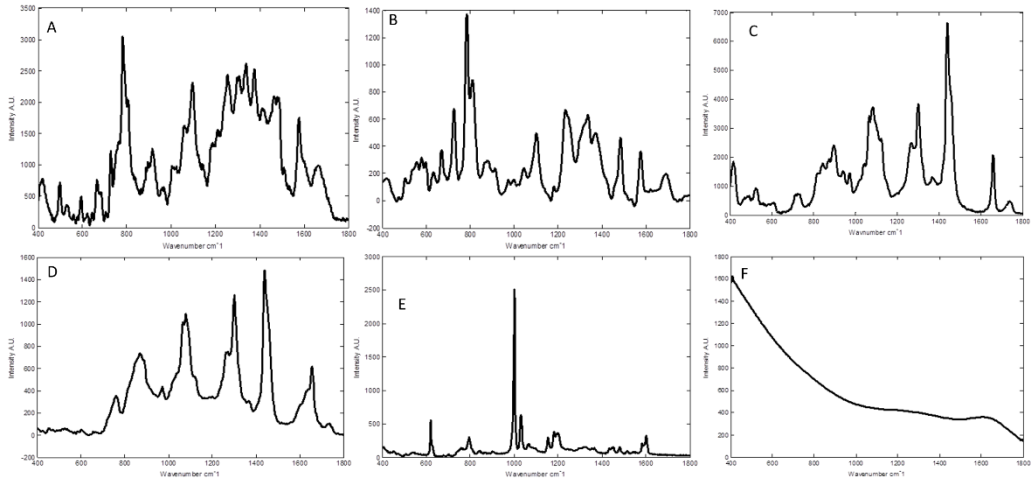


Figure 7.2: Clusters representing the Nucleus (A), Perinuclear 1 (B), Cytoplasm (C), Nucleolus (D), Perinuclear 2 (E) and Polystyrene nanoparticles (F).

Future work may also look to develop the algorithms of sePCA and SCCA, this could possibly look at areas such as disease diagnostics, coupling these methods with classifier algorithms such as SVM could lead to the development of novel approaches, with a high sensitivity and specificity for certain disease states.

While there are some limitations to this study, only Raman spectroscopy is used, possibly incorporation of other spectral techniques such as IR, CARS, SRS, might added to the scope of this study, expansion to include real data in the evaluations could progress the methods used and see them progress to routine usage in data analytics.

A sound knowledge of the workings of these and other multivariate statistical methodologies validated and verified in simulation, may lead to more robust and accurate multivariate statistical protocols in the biomedical spectroscopy field. Other modalities may also benefit, such as CARS and SRS,

as these techniques begin to gain ground and move towards a full spectrum video rate imaging technology [20,21].

Advanced simulations may also be used to identify possible artefacts from instrumental error and sample preparations, by the known introduction of these changes and their effect downstream at the analytical stage.

In conclusion advanced simulations have been demonstrated to shed light on multivariate statistical methodologies used for *in-vitro* screening for Raman spectroscopy. The knowledge demonstrated in this thesis may aid in the understanding and development of real data protocols to ensure accurate, valid and correct application of multivariate statistical methods in the future.

7.2 References

1 THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION, *Off. J. Eur. Union*, 2010, 33–79.

2 U. S. Congress, 2001, 2721–2725.

3 W. Russell, R. Burch and C. Hume, *The principles of humane experimental technique*, Methuen, London, 1959.

4 A. Tfayli, F. Bonnier, Z. Farhane, D. Libong, H. J. Byrne and A. Baillet-Guffroy, *Exp. Dermatol.*, 2014, **23**, 441–3.

5 F. Bonnier, M. Keating, T. Wróbel, K. Majzner, M. Baranska, A. Garcia, A. Blanco and H. J. Byrne, *Toxicol. Vitr.*, 2014, **29**, 124–131.

6 F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 322–32.

7 P. Bassan, A. Kohler, H. Martens, J. Lee, H. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke and P. Gardner, *Analyst*, 2010, **135**, 268–277.

8 H. Nawaz, F. Bonnier, A. D. Meade, F. M. Lyng and H. J. Byrne, *Analyst*, 2011, **136**, 2450–63.

9 H. Nawaz, A. Garcia, A. D. Meade, F. M. Lyng and H. J. Byrne, *Analyst*, 2013, **138**, 6177–84.

10 P. Knief, C. Clarke, E. Herzog, M. Davoren, F. M. Lyng, A. D. Meade and H. J. Byrne, *Analyst*, 2009, **134**, 1182–91.

11 J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers and H. J. Byrne, *Analyst*, 2012, **137**, 1111–9.

12 E. Efeoglu, M. Keating, J. Mcintyre, A. Casey and H. J. Byrne, *Anal. Methods*.

13 M. a Maher, P. C. Naha, S. P. Mukherjee and H. J. Byrne, *Toxicol. Vitr.*, 2014, **28**, 1449–60.

14 Organisation for Economic Co-Operation and Development, *ENV/JM/MONO*, 2013, **6**, 1–45.

15 C. Matthäus, T. Chernenko, J. a Newmark, C. M. Warner and M. Diem, *Biophys. J.*, 2007, **93**, 668–73.

16 M. Miljković, T. Chernenko, M. J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

17 F. Bonnier, a D. Meade, S. Merzha, P. Knief, K. Bhattacharya, F. M. Lyng and H. J. Byrne, *Analyst*, 2010, **135**, 1697–703.

18 K. Klein, A. M. Gigler, T. Aschenbrenner, R. Monetti, W. Bunk, F. Jamitzky, G. Morfill, R. W. Stark and J. Schlegel, *Biophys. J.*, 2012, **102**, 360–8.

19 F. Bonnier and H. J. Byrne, *Analyst*, 2012, **137**, 322–332.

20 N. L. Garrett, A. Lalatsa, I. Uchegbu, A. Schätzlein and J. Moger, *J. Biophotonics*, 2012, **5**, 458–68.

21 J. Moger, B. D. Johnston and C. R. Tyler, *Opt. Express*, 2008, **16**, 3408–19.

List of publications

Keating ME. and Byrne HJ. Seeded Principal Component Analysis for biochemical screening using vibrational spectroscopy. Submitted December 2015, *Analyst*.

E. Efeoglu, M. Keating, J. Mcintyre, A. Casey and H. J. Byrne, Determination of Nanoparticle Localisation within Subcellular Organelles in-vitro using Raman Spectroscopy Anal. Methods, 2015.

Byrne HJ, Knief P, Keating ME, Bonnier F. Spectral pre and post processing for infrared and Raman spectroscopy of biological tissues and cells. Chem Soc Rev. 2015 Oct 14.

Keating ME, Nawaz H, Bonnier F, Byrne HJ. Multivariate statistical methodologies applied in biomedical Raman spectroscopy: assessing the validity of partial least squares regression using simulated model datasets.2015. Analyst. 2015 Mar;16;140(7):2482-92. doi: 10.1039/c4an02167c.

Bonnier F, Keating ME, Wróbel TP, Majzner K, Baranska M, Garcia-Munoz A, Blanco A, Byrne HJ. Cell viability assessment using the Alamar blue assay: a comparison of 2D and 3D cell culture models. Toxicol In-Vitro. 2015 Feb;29(1):124-31. doi: 10.1016/j.tiv.2014.09.014.

Keating ME and Byrne HJ. Raman spectroscopy in nanomedicine: current status and future perspective. Nanomedicine (Lond). 2013 Aug;8(8):1335-51. doi: 10.2217/nnm.13.108.

Keating ME, Bonnier F, Byrne HJ. Spectral cross-correlation as a supervised approach for the analysis of complex Raman datasets: the case of nanoparticles in biological cells. Analyst. 2012 Dec 21;137(24):5792-802. doi: 10.1039/c2an36169h.

List of conference presentations

February 2012 Nano Impact Net, Dublin

June 2012 Biophotonics and Imaging Graduate Summer, Galway, Poster

November 2012 SPEC, Thailand, Poster

March 2013 Biopic, Dublin, and Poster

June 2013 Euro Nano Forum

November 2013 Chemometrics Ireland Meeting, Dublin, Talk

August 2014 SPEC 2014, Krakow, Poster, Junior Scientific Committee

April 2015 CLIRSPEC 2015, Exeter, Poster